A Deep Learning Pipeline for Comparing Microglia Morphology in Centenarians and Alzheimer's Disease Patients

Christopher Charlesworth[1]
Supervisor(s): Xucong Zhang[1], Maruelle Luimes[2]
[1]EEMCS, Delft University of Technology, The Netherlands
[2]Amsterdam University Medical Center, The Netherlands

TU Delft in collaboration with Amsterdam University Medical Center
A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Masters of Data Science and Artificial Intelligence Technology

Submitted on 2026-06-11
Defended on June 19, 2026

Name of the student: Christopher Charlesworth
Thesis committee: Xucong Zhang, Maruelle Luimes, Klaus Hildebrandt

1 Introduction

Alzheimer's Disease (AD) is one of the most common neuro-degenerative diseases and effects over 50 million people worldwide [8]. Although AD may appear to be an unavoidable consequence of aging, this is not always the case. Some individuals live to over 100 years old and are still cognitively healthy at the time of their death. These exceptional individuals are referred to as centenarians, and some have kindly chosen to participate in the 100 plus study lead by Amsterdam University Medical Center. Within the participating centenarians, 83% were independently assessed to be cognitively healthy, while over 50% were living fully independently, highlighting just how remarkable these individuals are [10]. By collecting and analyzing a variety of biological data, including post-mortem brain tissue and neuro-psychological assessments, the 100 Plus study aims to uncover the biological mechanisms that protect these individuals from cognitive decline [10]. The present thesis contributes to this goal by analyzing scans of centenarian brain tissue and AD patients, to identify differences in their microglia morphology that could be linked to longevity and AD. Such insights help us better understand Alzheimer's disease and may support the development of more effective treatments in the future.

Microglia are the resident immune cells of the Central Nervous System (CNS) and are critical for the maintenance of our brains [26]. They can exhibit a variety of cell states, some of which cause excessive neuroinflammation which has been linked to AD progression, in particular $A\beta$ and tau pathology [5]. For more information, see Section 2.1. These cell states are visually identifiable in stainings, as microglia change their shape (morphology) based on their functional state [28]. Therefore, by identifying and quantifying microglia morphologies, we can gain insights into what cellular activity was ongoing in a scan. By comparing the microglia morphologies of AD patients and centenarians, we can identify links between microglia cellular activity and AD progression or longevity. These links are critical for understanding both the role of microglia in AD and how they may have enabled centenarians to maintain cognitive health into extreme old age.

Performing this analysis is not straightforward, as existing methods fall short in several key ways. Firstly, manual labeling could be an approach but is prohibitively time-intensive and impractical for this work, as the total number of cells analyzed across the cohort exceeds one million. Furthermore, due to microglia's complex shape, human annotations can also be error prone and often differ across annotators [38]. Next, many semi-automated image processing methods have been employed to quantify microglia morphologies, such as custom ImageJ plugins. However, image processing alone generally leads to very poor performance as microglia are very diverse both in terms of morphology and staining intensity [38]. In recent years, computer vision techniques such as YOLO and U-Net have shown to achieve significantly better performance than image processing methods [2]. Despite this potential, previous work has typically focused on isolated stages of the analysis, such as identification or segmentation alone. Only one other study has attempted a full pipeline, yet

it demonstrated poor performance and relied on a rudimentary evaluation [11]. See Section 3 for more information.

This work aims to address these limitations by developing an end-to-end machine learning pipeline incorporating all necessary steps for microglia morphology analysis. This includes microglia detection, segmentation, morphological feature extraction, dimensionality reduction, clustering to identify cell states and finally statistical testing between AD patients and centenarians. Beyond the development of a comprehensive end-to-end pipeline, the primary novel contributions of this work include:

- **Rat-to-human transfer learning:** Developing a method to improve microglia detection performance on humans by first learning from more available rat data.
- **Morphological diversity active learning:** Developing a novel active learning strategy that is specific to microglia by prioritizing new labels that are morphologically different from the current training data.
- **Topology-aware segmentation:** Implementing topological loss components and custom data augmentations that enable segmentation models to better preserve the fine branching nature of microglia.
- **Morphology-centric evaluation:** Defining a novel segmentation metric that quantifies model performance based on its ability to predict morphological features.
- **Spectrum-based soft clustering:** By incorporating soft cluster assignment, we were better able to model the continuous spectrum of microglia morphologies rather than assigning discrete states.
- **Robust statistical framework:** By applying multiple scan-level statistical tests, including Dirichlet, Mixed-Effects models and PERMANOVA, we were able to identify differences in microglia morphology between the diagnosis groups while accounting for intra-patient correlation and covariates.

Building upon these contributions, this research aims to answer the following specific research questions, categorized by pipeline component:

- **Cell Detection and Data Efficiency**
  - Can cross-species transfer learning from rats to humans improve microglia detection performance compared to only using human labeled data?
  - Can active learning techniques be used to improve microglia detection performance beyond that achieved by additional random training data? If so, which technique is most suitable between edge confidence selection and prioritizing morphological diversity?

- **Segmentation Quality**
  - Does a custom topology-aware U-Net outperform the foundational model Segment Anything Model (SAM) on accurately segmenting microglia's complex shape?
  - Do domain-specific augmentations and topology-aware loss functions (clDice, Betti) improve the preservation of fine structural details compared to standard training approaches?
- **Morphological Representation**
  - Does the linear dimenionsality reduction technique Principle Component Analysis (PCA) better capture morphological features than a non linear approach like t-SNE?
  - Which soft and hard clustering methods are most appropriate for capturing microglia morphologies?
- **Biological Characterization**

  Note: The Diagnosis groups referenced below are Alzheimer's disease patients, centenarians without AD pathology, and centenarians with AD pathology. See section 6.5 for more information.
  - Are there differences in individual morphological features (e.g., skeleton length, soma area) between the diagnosis groups after adjusting for covariates like sex?
  - Are there differences between the diagnosis groups when comparing the multivariate morphological profiles between the diagnosis groups?
  - Does the distribution and variance of morphological states (clusters) differ between the diagnosis groups?

The remainder of this work is structured as follows. Section 2 provides background information on the relevant biology, computer vision, and clustering techniques applied. Following this, Section 3 discusses related work and its limitations, broken down by pipeline step. Subsequently, Section 4 explains the unique dataset and annotations used for this thesis. From there, Section 5 covers the methodology of the entire pipeline, including the active learning techniques employed. Next, Section 6 presents all experiments conducted as well as the evaluation metrics. Building on this, Section 7 discusses the results including the statistical analysis comparing the diagnosis groups. Next, Section 8 summarizes the main insights, highlights limitations, and suggests directions for future work. Finally, Section 9 includes a reflection on the broader scientific landscape and societal context of this thesis, while Section 10 discusses the use of generative AI in its creation.

2 Background

2.1 Biology Background

2.1.1 Microglia

Microglia are the resident immune cells of the CNS, constantly surveying the brain environment for pathogens, damage signals, and changes in the extracellular microenvironment [26]. Upon detecting a problem, microglia carry out a clearance function known as phagocytosis, in which they engulf and remove dead cells or cellular debris [37]. Furthermore, microglia help regulate neuroinflammation through the release of cytokines, a process that can be beneficial in some contexts but, when chronic, may become harmful and contribute to neurodegenerative diseases such as AD [27]. Beyond these immune functions, microglia also play a critical role in normal brain development through synaptic pruning, a process in which excess synapses are removed [25]. Moreover, microglia are abundant throughout the brain, making up approximately 10–15% of all brain cells [19], though their distribution varies across regions. For example, their abundance differs between white matter and gray matter, and their morphology is likewise shaped by the regional environment [16].

Microglia morphology is the visual appearance of microglia that often reflects their functional state [26]. Previous research has identified a number of morphologies that represent different functional states including Homeostatic, Reactive, Amoeboid, Hyper-ramified and Rod [28]. See Figure 1 for examples of what these morphologies look like, both as masks and illustrations, that make their differences more clear. Many of these morphologies are characterized by the microglia soma and processes; the soma is the circular body located near the center of the cell, while the processes are the arm-like extensions. Below is an overview of each morphology based on the classifications described in [28].

- **Homeostatic:** This is the most common microglia morphology and is also commonly referred to as Surveillant, Ramified or Non-active. Cells in this state typically have a smaller soma with several branching processes that aid the cell in identifying signs of infection or dead cells.
- **Reactive:** Once a microglial cell detects a stimulus with which it will interact, it enters a Reactive morphology. This state is generally characterized by fewer but thicker branches, and the soma may also increase slightly in size. Because it is visually quite similar to the homeostatic morphology, many computer vision based studies group these two states together. However, Homoeostatic and Reactive morphologies represent different functional states within the cell, as Reactive morphologies produce inflammatory cytokines and begin to phagocytose cellular debris.
- **Amoeboid:** Amoeboid microglia are commonly associated with active phagocytosis, although morphology alone does not definitively confirm

that phagocytosis is occurring. Many studies also refer to this morphology as Fully Activated or Phagocytic.

- **Hyper-ramified:** This morphology is typically seen as a transition state between Homeostatic and Reactive with the cell beginning to retract its processes.
- **Rod:** These microglia are uniquely identifiable by their long, skinny somas as well as relatively few elongated processes. This morphology is less common than others but frequently appear near damaged neurons, for example after an injury or a stroke.

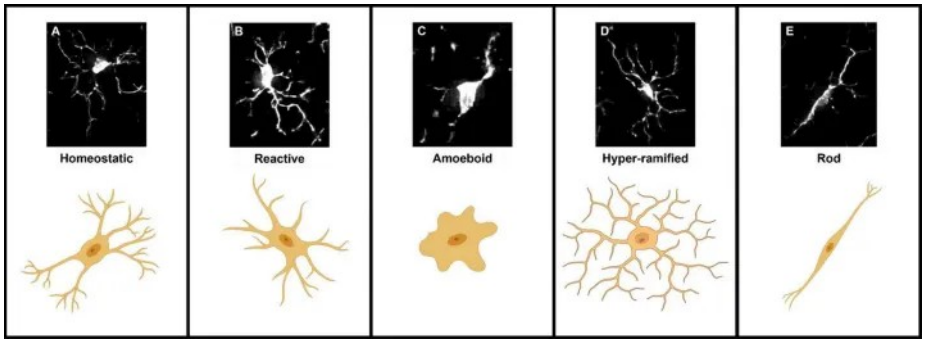


Figure 1: Example illustrations and masks of the main microglia morphologies, including **Homeostatic**, which have many processes; **Reactive**, which have fewer but thicker processes; **Amoeboid**, which have no processes and represent an activated state; **Hyper-ramified**, which has a very complex and branching process structure; and finally **Rod**, which have long, thin somas with typically no processes. Figure originally from [28].

It is important to note that these morphologies are not discrete categories, but instead lie along a spectrum of cellular states. This is because microglia undergo gradual and context-dependent structural changes, meaning that intermediate forms often exist between commonly described morphological classes [4].

2.1.2 Alzheimer's Disease

AD is one of the most common forms of dementia and is commonly identified by amyloid-beta deposition and intracellular hyperphosphorylated tau accumulation [21]. **Amyloid-beta,** $A\beta$, is a small protein fragment that can aggregate together outside of neurons to form amyloid plaques, which is one of the hallmarks of AD. There are multiple forms of $A\beta$ with $A\beta 42$ being most associated with AD. Moreover, amyloid plaques are not the only type of plaques linked to AD, with neuritic plaques also being one of the primary criteria for AD diagnosis.

Tau is a protein found inside neurons that is used to stabilize microtubes that support the structure of the cell. Neurofibrillary tangles can occur when bundles of twisted tau, also known as hyperphosphorylated tau, aggregate inside a neuron, affecting the neurons ability to function properly. This has been strongly linked to the neurodegeneration seen in AD [21].

As microglia are responsible for detecting and removing any pathogens within the brain, they identify the $A\beta$ plaques and become activated with the goal of containing or clearing the amyloid material. If this activation happens too frequently or incorrectly, it can lead neuroinflammation or neuron damage [21].

2.2 Background Computer Vision

YOLO

Object detection is the task of identifying and locating objects within an image, typically by drawing bounding boxes around them. There have been many different techniques developed to achieve this, including YOLO which was introduced by Redmon et al in [29]. YOLO differs from many of its predecessor models by treating object detection as a single problem rather than splitting it into multiple stages like R-CNN and Fast R-CNN. YOLO works by dividing an image into an S * S grid, where each cell predicts bounding boxes, confidence scores, and class probabilities for any objects it contains. This is all performed in a single forward pass through a Convolutional Neural Network (CNN). Next, the final detections are filtered using Non-Maximum-Suppression (NMS) to only keep the strongest boxes and remove overlapping duplicates [29]. YOLO performs strongly across a variety of domains because it uses the full image context when making predictions, which helps reduce background false positives. It is also end-to-end trainable, so all components are jointly optimized for detection, unlike two-stage detectors such as R-CNN [29].

U-Net

Image segmentation is the process of dividing an image into meaningful regions by assigning a class label to each pixel. U-Net is a common segmentation architecture developed by Ronneberger in [30] for the purpose of segmenting biomedical images with small datasets. It is characterized by its encoder-decoder 'U' shape, with the encoder learning what is in the image and the decoder learning where the objects are. Skip connections are a key feature of the U-Net architecture, acting as direct links from earlier layers to later layers, bypassing the intermediate layers. They help preserve low-level spatial information, such as processes in microglia, by sending high-resolution features to the decoder so that it can identify object boundaries more precisely [30]. Like YOLO, U-Net is also end-to-end trainable, meaning there is no need for careful hyperparameter tuning as seen in traditional image processing based segmentation architectures. Furthermore, U-Net works incredibly well on small datasets due to the skip connections and overall architecture being highly-data efficient. Finally, U-Net's performance also significantly improves when data augmentations are used to artificially expand the annotated dataset [30].

AD is an incredibly common disease with it effecting over 50 million people world wide and commonly being associated with alpha beta tau accumulation.

2.3 Clustering Background

2.3.1 Hard Clustering

KMeans
KMeans is one of the most commonly used clustering algorithms as it is highly applicable to a variety of domains. The goal of KMeans is to minimize the distances between points within the same cluster [13]. It does so by selecting random points to be the starting centroids of the clusters. Then, each data point is iteratively assigned to a cluster based on it being closest to that centroid. The centroids are then updated as the mean of all points in the cluster with this process repeating until all points have been assigned to a cluster [13].

Agglomerative clustering
Agglomerative clustering is a bottom up approach where each data point is initially treated as its own cluster. Next, the algorithm iteratively merges the two closest clusters until you have reached the target number of clusters [43]. This naturally produces a hierarchical relationship between the clusters which can be useful for identifying subtypes in microglia morphologies [42].

DBSCAN
DBSCAN is a density based clustering method, meaning it separates clusters based on there being a region with low density between them rather than a large distance like KMeans. It does this by grouping together points that have enough nearby neighbors within a chosen radius, while labeling isolated points as noise [43].

2.3.2 Soft Clustering

Fuzzy C-Means
Fuzzy C-Means is a soft clustering extension of KMeans, meaning the clusters often look similar. In hard clustering, each point belongs to exactly one cluster. However, soft clustering assigns each point a membership value for every cluster, ranging from 0 to 1, where the values across all clusters sum to 1. The Fuzzy C-Means algorithm initializes its clusters based on those that KMeans identifies and then calculates the cluster memberships based on a point's distances to each cluster's centroid. Next, the centroids are updated based on taking a weighted average of the data point's positions, where the weights are the cluster membership values. After the centroids are updated, the membership values are recalculated. This process repeats until the memberships or centroids change very little [43].

Gaussian Mixture Model (GMM)
GMM is a soft clustering method similar to Fuzzy C-Means, except it calculates the cluster proportions differently. It treats each cluster as a Gaussian component with its own mean, covariance matrix and mixture weight. Next, for each data point, it calculates its cluster proportions based on the probabilities modeled in these gaussians and then normalizes the probabilities to add up to 1 [6]. This means it is more likely to assign harder cluster proportions than Fuzzy C-Means, as the posterior probabilities tend to concentrate more strongly around the most likely Gaussian component.

3 Related Work

3.1 Microglia Identification

Past studies have used traditional image processing techniques to both segment and identify microglia in a variety of scanning technologies. For example, in 2012 the authors of [41] automated the counting of rat microglia and reported an accuracy of more than 80% compared to manual labeling. They achieved this by designing a custom multistage image-processing pipeline with steps like boosting, intensity outlier normalization, global thresholding and morphological filtering. However, traditional image processing methods have critical flaws compared to deep learning techniques, mainly in their reliance on manual hyperparameter tuning and thus suboptimal performance and poor generalization to other datasets.

In recent years, there has been significant performance improvements in microglia detection and quantification through the use of CNNs like YOLO. For instance, in [38] the authors compared their custom YOLO v3 model to image processing techniques like plugins from ImageJ, the semi automated tool Ilastik as well as manual annotations from humans. The result was a significant performance improvement compared to previous methods, with an F1-score of 0.92 compared to ImageJ's 0.75. Interestingly, the authors also found that the human annotations were imperfect by discovering high inter-annotator variability. This means that two different annotators would often give different annotations for the same image, due to them making errors and some microglia being ambiguous. Therefore, they also compared their model's performance to the average annotator's performance and found that their YOLO model was more accurate than the humans, while also being 170 times faster.

Similarly, in [2] the authors performed a robust comparison of many deep learning architectures vs traditional methods for microglia quantification in rats, mice and non-human primates. In rats, they found that YOLO had the highest correlation with their manual annotations with a Pearson R2 value of 0.942 compared to Faster R-CNN at 0.919 and RetinaNet at 0.935. Overall, the deep learning methods significantly outperformed the image processing techniques of Ilastik and Fiji with R2 values of 0.823 and 0.621 respectively. These findings were also consistent across species and indicate that YOLO is likely the

strongest available deep learning architecture for microglia identification.

There are several limitations present in previous literature that have been addressed in this work. First of all, since image-processing methods rely heavily on manual hyperparameter tuning, we have instead opted for a fully machine learning approach for our detection. Furthermore, previous YOLO microglia detection methods use their training data very inefficiently, resulting in them needing a very large training set to achieve strong performance. This means labeling their dataset is very time consuming and logistically difficult as it should be done by an expert. For example in [11] they annotated over 88000 cells and other papers frequently reach over 10000 [1]. This work aims to address this limitation by achieving strong results with a much smaller dataset through the use of active learning strategies. Additionally, prior work was limited to single-species training data, whereas we leverage inter-species transfer learning to improve detection performance without requiring additional annotations.

3.2 Microglia Segmentation

Just as with detection, many early work in microglia segmentation used image processing pipelines. For example, the authors of [7] developed an image segmentation pipeline with many custom steps specific to 2D segmentation of microglia cells near $A\beta$ plaques. These custom steps included connecting disconnected fragments of processes, removing cells that do not meet certain morphological properties eg. no soma and enhancing the branches before segmentation to ensure they were not removed. Although this paper did find morphological differences between microglia near $A\beta$ plaques, the overall evaluation of their pipeline lacks rigor. Like many other image-processing pipelines [42], the authors of this study evaluated performance only qualitatively, meaning their findings may not be robust. The authors make no mention of performing any quantitative evaluation of their segmentation using metrics such as Dice.

Recently, many papers have applied U-Net to segment microscopy images, with it frequently outperforming image processing techniques as well as other baselines such as Segment Anything Model (SAM) [40]. AIStain, released in 2025, includes U-Net models to segment microglia in bright-field images before staining [44]. The authors found that their custom U-Net drastically outperformed the industry standard Cellpose 3 and SAM 2 with an AUC ROC of 0.964. However, they did not evaluate whether the morphological features of the cells were correctly predicted by their U-Net, meaning it is possible that some thin processes are lost. Overall, this work showed that U-Net's architecture is well suited for microglia segmentation with its complex processes and diverse morphologies. Although these models were trained directly on bright-field scans rather than Iba1 (the focus of this work), other research has shown that these findings are applicable to Iba1 scans [44].

TrAIce3D is perhaps the most advanced segmentation model for 3D microglia scans, with it both segmenting the soma and branching processes using a 2 stage pipeline [1]. The authors found that their transformer-based U-Net outperforms CellPose by over 0.2 in F1 score for soma segmentation. Importantly, the authors found that their segmentation performance on somas and processes (branches) was significantly biased, with the soma being significantly easier for their model to segment. To further quantify their process segmentation quality, they calculated the average path length difference and Hausdorff distance between their predicted masks and the target masks. Although this paper focused on 3D segmentation while we are using 2D data, its evaluation methodology indicates a strong need for a more robust microglia segmentation evaluation than just traditional pixel based metrics like Dice, F1, etc.

Although the field of microglia segmentation has improved significantly in the last few years, there are still many significant limitations present in the previous literature that this work aims to address. Firstly, regarding evaluation, many image-processing methods were only evaluated qualitatively while later work mostly focused on traditional segmentation metrics like Dice and IoU. These are insufficient for microglia segmentation as they do not capture whether morphological features were accurately predicted for downstream tasks. This work aims to address this limitation by introducing a custom morphology score based on 25 morphological features. Next, existing segmentation studies often lacked microglia-specific augmentations, topology-aware losses, and targeted postprocessing. By contrast, this work aims to develop a segmentation model truly designed around microglia and their unique morphology with several custom components that significantly improve performance.

3.3 Clustering

As discussed in 2.1.1, microglia can exhibit different morphological states like Reactive, Homeostatic, Amoeboid etc. Rather than using labels to identify these morphological groups, some research has instead opted for an unsupervised approach. For example, the authors in [42] used hierarchical clustering to identify microglia morphologies that represent cell states. Instead of clustering directly on the morphological features, they first applied Principle Component Analysis (PCA) to reduce the dimensionality and then they clustered on this latent space. This approach was taken because many of their 27 morphological features were correlated with one another and represented only a few key ideas. While the authors identified regional heterogeneity in the brain, they did not identify sexual dimorphism in rats.

There are several shortcomings present in both the paper described above and in microglia clustering research more generally. For example, they only assess one clustering method and if they do employ dimensionality reduction, they only use PCA. Contrastingly, this work compares and assesses multiple dimensionality reduction techniques and clustering methods to determine which is actually most suitable for identifying microglia morphologies. Furthermore, earlier studies handled morphology grouping very rigidly, even though microglia are known to lie on a continuous morphological spectrum [4]. Our analysis addresses this issue by representing microglia morphologies with soft clusters that better capture these transitional microglia states.

3.4 Full Pipelines

One key limitation of almost all deep learning papers in the microglia field is that they apply machine learning only to one small part of the analysis, for example, just cell detection. In 2025 a new paper [11] instead presented an end-to-end deep learning pipeline for analyzing microglia morphology in rats. Their goal was to assess how hypothermia and cardiac arrest influenced microglia morphologies in different regions of the brain. To do so, they implemented YOLO to identify microglia, a U-Net model for segmentation and finally a supervised classifier for determining a cell's morphological group. While their pipeline did combine multiple stages together, its performance is lacking, with a mAP50 of 0.796 for detection and a Dice score of 0.807 for segmentation. Overall, this poor performance can be attributed to a lack of custom microglia specific steps in their pipeline with it being very task agnostic. In contrast, this thesis aims to improve performance significantly by using custom microglia components like data augmentations, topology-informed loss components, postprocessing segmentation masks, inter-species transfer learning and active learning.

4 Data

4.1 Scan Data

The dataset used in this work is split into two main groups: (1) patients who were diagnosed with AD prior to death and (2) centenarians, people who lived to be more than 100 years old. In total, 91 individuals were analyzed in this work, 84 of which were centenarians and 7 had AD, meaning there is a significant class imbalance. However, when the brain tissue of centenarians was analyzed postmortem, some individuals exhibited high levels of AD pathologies even though they were assessed as being cognitively healthy and lived for an extraordinary amount of time. These individuals should be treated as a distinct group from the other centenarians as they were actively combating AD at the time of their death. To determine which centenarians had sufficient AD pathology to be placed in this group, we used the NIA-AA ABC framework described in [12]. This framework combines the amyloid-beta plaque distribution, neurofibrillary tangle stage, and neuritic place score to classify individuals into 'Not', 'Low', 'Intermediate' or 'High' levels of AD neuropathologic change. See Section 2.1 for more information on what these terms indicate. Centenarians classified as 'Intermediate' or 'High' were assigned to the '100+_with_pathology' group, whereas all remaining centenarians were assigned to the '100+_without_pathology' group. In total there were 39 individuals placed in the 100+_with_pathology group while 45 belonged to the 100+_without_pathology group. Based on these three diagnosis groups, multiple statistical tests were conducted to detect morphological differences between them, both at the cluster level and on the raw morphological features.

The scans themselves were obtained postmortem from individuals who donated their brain to the Netherlands Brain Bank (NBB) and the dataset was provided by the 100-Plus group, situated at Amsterdam University Medical Center (AUMC). Iba1 immunohistochemistry was used to visualize the microglia in the temporal cortex by using an antibody that binds to the Iba1 protein that is generally present within microglia. After binding, a chromogenic reaction makes the microglia visible under a microscope as a brown color (see Figure 2). The stained brain tissue was digitalized using a automated slide scanner, resulting in 1 digital scan per participant. Only the gray matter was analyzed in this study, as microglia in gray and white matter exhibit distinct morphological characteristics [16] and the main AD hallmarks of $A\beta$ plaques and neurofibrillary tangles are mainly found in gray matter. Annotations for white vs gray matter were also provided per scan by AUMC.

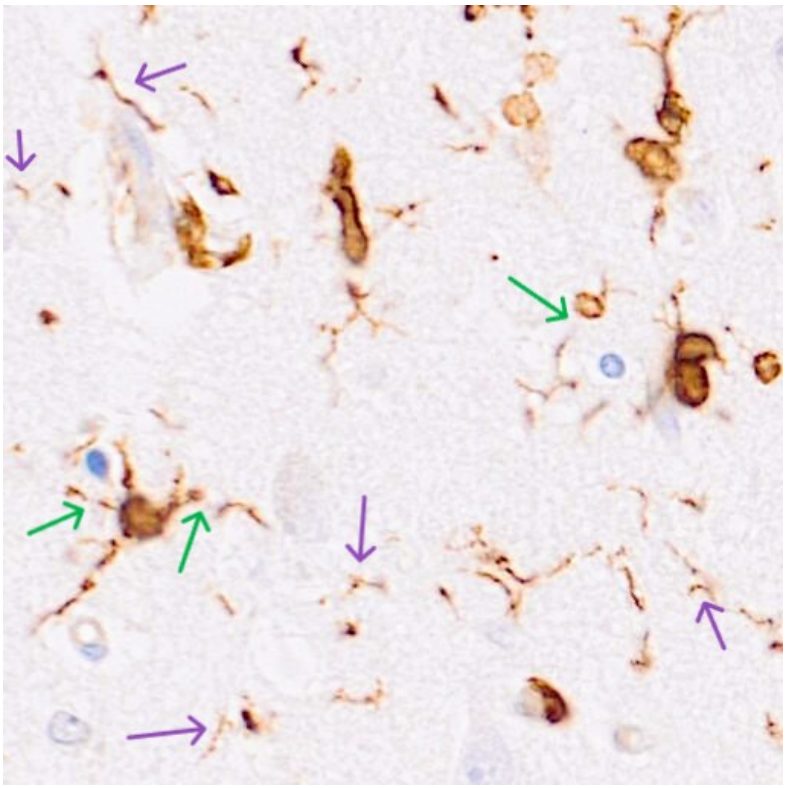


Figure 2: Example 512×512 pixel tile of Iba1 stained human microglia from a centenarian. The arrows show examples of the common data quality issues caused by taking 2D scans of 3D microglia. Purple arrows show some of the fragments of cells that should not be analyzed, while the green arrows indicate disconnected components of the same cell caused by the scan cutting the branch.

It is important to note that the data analyzed in this work consist of 2D slices of microglia, while the cells themselves are 3D structures. Figure 3 shows an illustration of how a single microglia cell can be split over multiple slices. This can cause small fragments that appear between microglia, for example those indicated by the purple arrows in Figure 2. These fragments are parts of other microglia that are not fully captured in the 2D slice and therefore should not be analyzed. Conversely, it is also possible for slightly disconnected processes to belong to the same microglia cell, but appear separated because they curve out of the current 2D slice; for example, those indicated with green arrows in Figure 2.

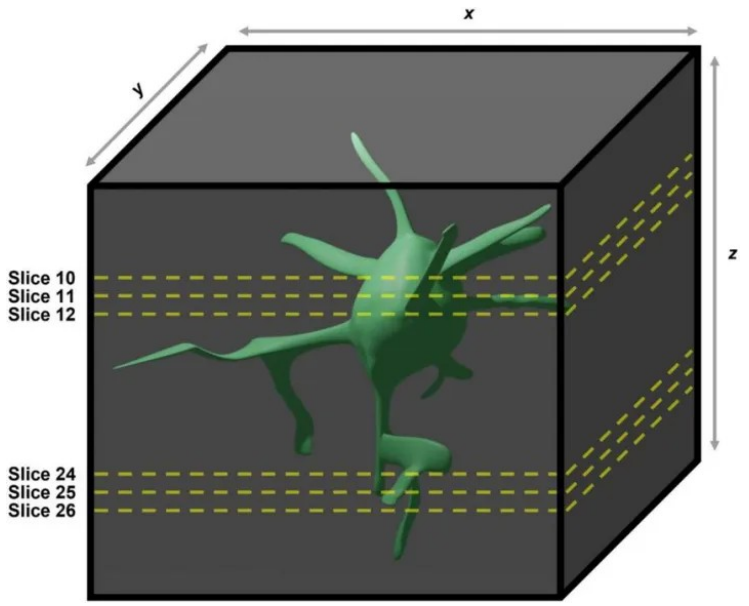
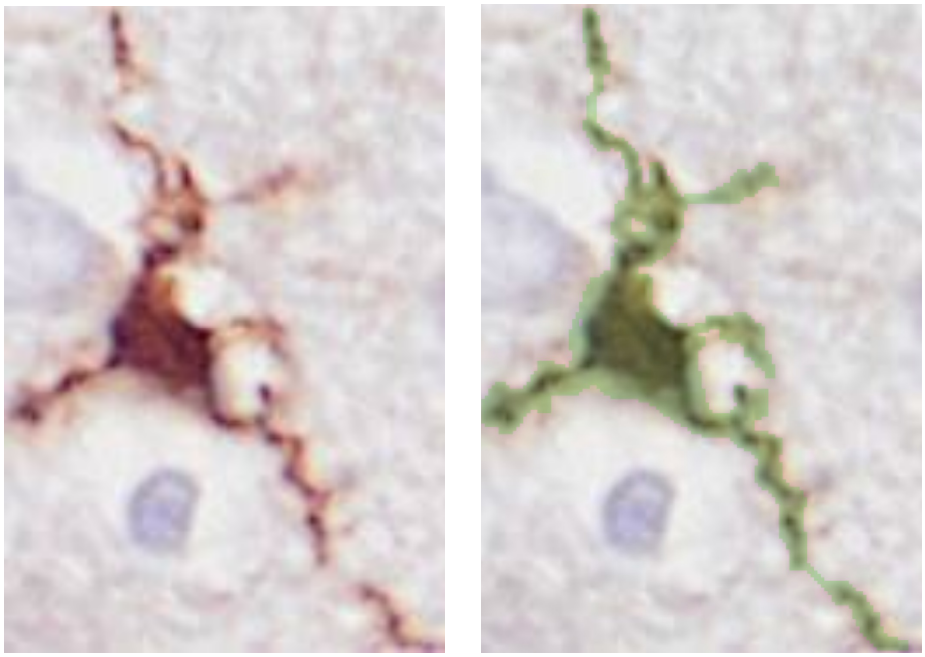


Figure 3: Visualization of how 2D slices taken from a 3D microglia can result in fragments of cells being stained. Figure originally introduced in [28]

.

4.2 Annotations and Preprocessing

With the individual scans being incredibly large files at over 100,000 pixels by 200,000 pixels, it was infeasible to load them directly into memory for training or inference. Thus, the scans were split into 512 by 512 tiles using a sliding window that produces thousands of tiles per scan. Initially, a total of 1009 tiles were annotated for identification totaling 4455 individual microglia. This dataset was later expanded using multiple active learning methods, see section 5.2 for more information. For segmentation, 1009 segmentation masks of individual cells were manually annotated using QPath and adjusted using CVAT, see Figure 4 for an example of these annotations.



(a) Microglia cell (b) Segmentation mask overlayed

Figure 4: Example of an individual microglia cell with a manually labeled segmentation mask used for training the U-Net.

4.3 Rat Data

To improve detection performance and make better use of the relatively small labeled human dataset, we incorporated rat microglia images as additional training data. Rat microglia share some features with human microglia, including a soma and branching processes. Furthermore, rat datasets are more readily available online because rat tissue is more accessible for analysis. The exact rat dataset we used was a subset of the cells labeled and introduced in [2]. Only 624 tiles were selected as the other data came from different staining methods and thus did not improve the performance on our dataset. Figure 5 shows an example of rat microglia stained with Iba1. Rather than mixing the rat and human data within the same training set, we instead chose to first pretrain on the rat data and then fine tune on the human dataset. This involved first training for 100 epochs on the rat data and then a further 100 epochs on the human data but with a much smaller learning rate. Overall, this approach led to a significant performance improvement compared to only training on human data, see Section 7.1 for more information.

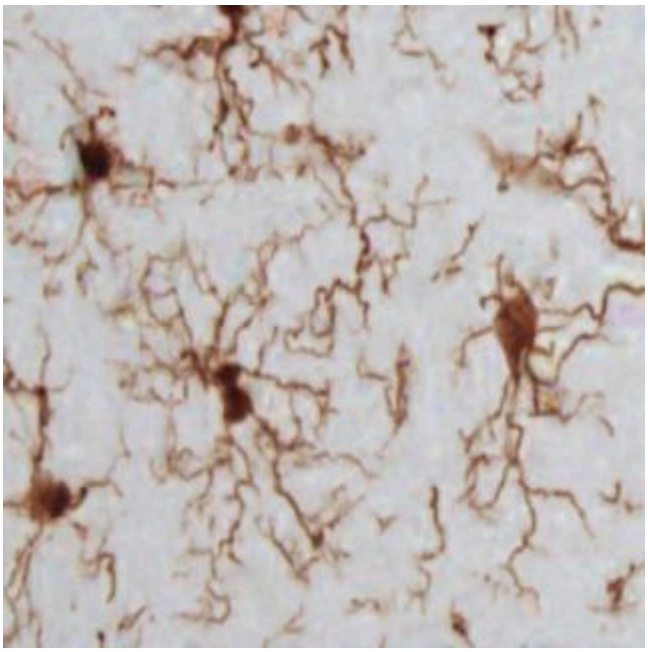


Figure 5: Example 512×512 tile of rat microglia used for pretraining our detection model. The dataset itself was originally introduced in [2] and was stained using Iba1.

5 Methodology

The main contribution of this work is the development of a fully automated machine learning pipeline for microglia morphology analysis. As the pipeline comprises multiple sequential stages, this methodology section is organized into a series of subsections, each addressing a specific component of the overall framework. The stages of the pipeline can bee seen in Figure 6 and they interact with one another as follows. First, the scans are split into 512×512 pixel tiles before the cells are detected using a YOLO model. The resulting bounding boxes are then passed into a U-Net segmentation model to separate the individual cells from the background. Based on the resulting masks, 25 morphological features are calculated that represent the overall shape of the cells. Next, we used clustering to identify the different morphologies present in the microglia. However, we do not cluster directly on the morphological features, but instead use dimensionality reduction to reduce the impact of many features being correlated with one another. Finally, we conducted a rigorous statistical analysis to determine whether there are differences in microglia morphology between the diagnosis groups. This was in terms of cluster proportions, individual morphological features and combinations of morphological features.

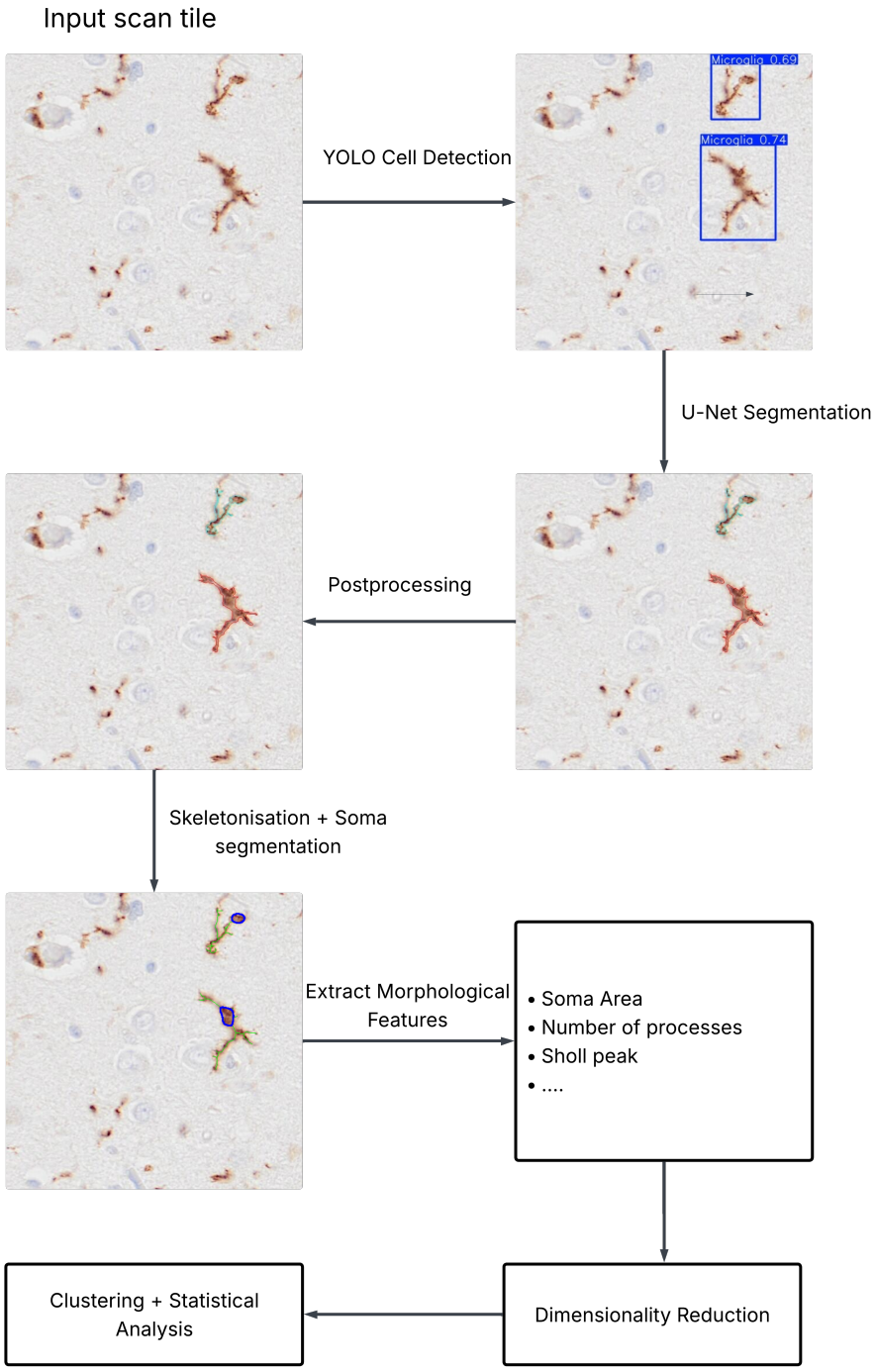


Figure 6: Diagram showing the input and outputs of all stages of the pipeline for an example 512×512 pixel tile. **1.** YOLO detects microglia and produces bounding boxes. **2.** U-Net segments each detected cell within the bounding box. **3.** Post-processing connects the different components of each cell. **4.** Morphological features are calculated from the cell mask, soma mask, and skeletonisation of the cell. **5.** Clustering identifies cell morphologies and diagnosis groups are compared via statistical analysis.

5.1 Cell Detection

The first stage in our pipeline is to find all of the microglia in the scans. This detection was done by training a custom YOLOv8 using Ultralytics to predict one class: microglia. The input to the model is a 512×512 pixel tile rather than the entire scan at once, due to the memory constraints discussed in 4.2 and because YOLO generally performs poorly on small objects [29]. By cropping the scans into tiles, the microglia appeared larger, making it easier to identify them. The resulting output from our YOLO model is a series of bounding boxes with associated confidence scores. NMS was then applied to remove duplicates before retaining only those with a confidence greater than 0.5 as true microglia. The model itself was trained for 100 epochs, with early stopping incorporated to prevent it from overfitting on the training data. A patience of 10 epochs was used, meaning that if no improvement was observed on the validation set for 10 epochs, training was stopped and the best-performing model was evaluated on the test set. To guide this training, a composite detection loss was used, consisting of binary cross-entropy for classification, Complete IoU for bounding-box regression, and Distribution Focal Loss for improved localization precision.

5.2 Active Learning

Active learning is a machine learning framework in which a model selectively chooses the most informative unlabeled samples for annotation in order to improve performance with fewer labeled examples [32]. This is the first study to implement this technique in the field of microglia detection, extending on its usefulness in a variety of domains [18].

This work has compared three active learning techniques to determine which is most appropriate for microglia detection. Firstly, cells were randomly selected to see what performance gains could be achieved by merely having a larger dataset. This random dataset can also be seen as a baseline, as no logic was used to select the new data points. Next, we used the confidence values outputted by our YOLO model as an indication for which cells had the highest uncertainty. We then annotated cells that had a confidence value between 0.45 and 0.55, i.e. cells near the decision boundary, with goal of providing informative samples to the model. One can understand this approach as a teacher giving a student extra practice exercises targeting the areas they were previously unsure about, thus helping them perform better on the next test. Finally, we selected new cells to annotate based on them being the most morphologically different to our current dataset. This was done by first running the entire pipeline: cell detection, segmentation, morphological feature extraction, and dimensionality reduction. This was followed by identifying which cells were furthest from the current annotations in the latent space. Figure 7 shows a visualization of this with all cells plotted on the t-SNE components that best capture the variance in their morphology. The black points represent the currently labeled data. As shown, there are many areas where the density is sparse, indicating that the current annotations are not fully representative of all possible morphologies. The colored

points represent the cells selected for annotation; a low rank (yellow) indicates a large distance from any currently labeled data point meaning they are the most informative. It is important to note that new cells were selected using a greedy strategy, in which previously selected cells were also treated as labeled when calculating distances. An analogy for this active learning strategy is a teacher giving students questions that are on the syllabus but very different from those seen in previous homework or quizzes, ensuring they are well-prepared for their final examination.

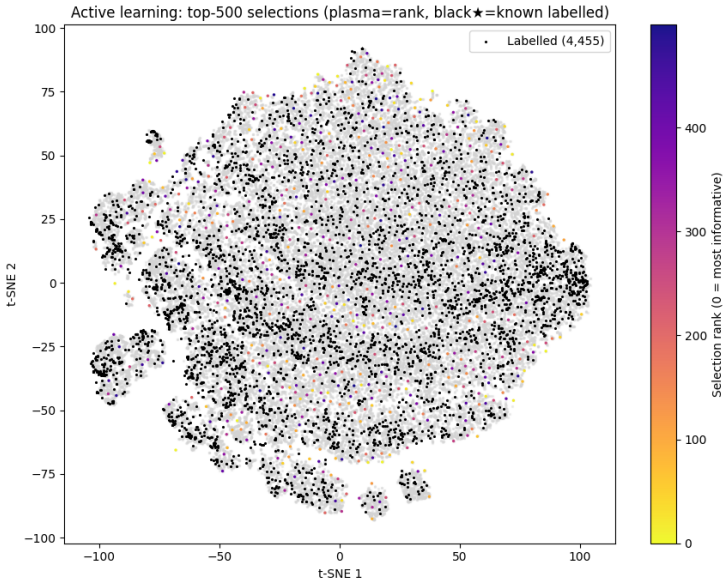


Figure 7: Plot of the 500 cells most morphologically distinct from the current labeled dataset, based on distances in the t-SNE embedding of morphological features. Black points show the initial dataset while the colored points are those selected using the morphologically diverse criteria, with a low rank indicating very informative data points.

One caveat of the current active learning implementation is that the YOLO model was trained on 512 x 512 tiles containing multiple cells, for which it predicts bounding boxes. Consequently, when annotating a new cell, all other cells within the same tile must also be annotated to avoid introducing false negatives into the training data. As a result, the confidence and morphology-based datasets also contain many cells that were effectively selected at random. This means some random noise could also be present in the results comparing these active learning method's performance.

5.3 Segmentation

5.3.1 U-Net architecture

Based on the bounding boxes identified in the previous section, we needed to determine what pixels were part of a cell and which were not. We did so with a custom U-Net model that performs whole-cell segmentation, producing a single mask for both the soma and processes. The architecture is a Convolutional Neural Network (CNN), meaning it is robust to local spatial variation and is well suited to learning hierarchical image features such as edges, textures and cell structures [20]. The model architecture can be seen in Figure 8 with it having 6 layers: 3 downsampling layers in the encoder and 3 upsampling layers in the decoder. The encoder is responsible for capturing contextual information while the decoder restores the spatial resolution such that the final mask matches the input size of the image. Max-pooling was used in the encoder for downsampling to progressively reduce spatial resolution and increase the receptive field, while bilinear interpolation was used in the decoder for upsampling. This compact setup was chosen instead of the deeper original U-Net to reduce model complexity and the risk of overfitting, which is especially important given the small size of our labeled dataset. Skip connections were also used in order to improve its performance on fine grained details like the microglia processes. Overall, while the model architecture is a relatively standard U-Net setup, several key microglia-specific modifications were made to the loss function, data augmentation, and evaluation procedure to ensure strong performance on microglia segmentation.

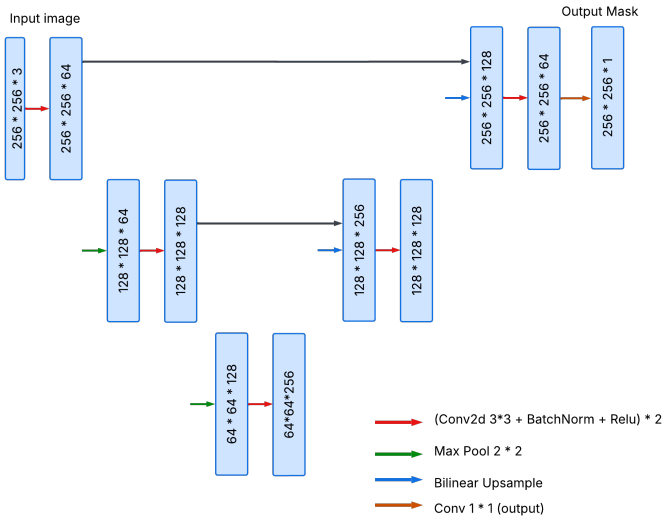


Figure 8: Architecture of the U-Net model used for microglia segmentation. Each block represents a different layer in the in the model architecture with the final block producing the predicted segmentation mask.

5.3.2 Loss Function

Multiple loss components were used in combination to achieve the best segmentation performance. Each component enabled the model to learn different aspects of an optimal microglia segmentation.

Binary Cross Entropy
Binary Cross Entropy (BCE) is one of the most commonly used loss functions in machine learning and is defined by the formula below, where N is the total number of pixels; y_i represents the ground-truth binary label for pixel i; and $\hat{y}_i$ represents the predicted probability for that pixel.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

This penalizes the model when its predicted mask probabilities are significantly different from the ground truth mask. This enables the model to learn general information about each pixel's probability, but does not penalize it for making morphological errors like disconnected components or missing processes.

ClDice
ClDice works in a very different manner than BCE as it aims to penalize the model when it makes topological errors, eg. missing a branch or including an extra one. It was first introduced in [34] with the goal of improving blood vessel segmentation performance. The authors found that two very similar segmentations can have the same Dice or BCE score, even though one preserves the topological features much better. This means that models performing well on traditional metrics may still fail to properly preserve topological features, such as branches. When analyzing microglia morphologies it is of the upmost importance that these topological features are preserved, therefore ClDice was implemented as an additional loss component.

ClDice works by comparing the skeletonized versions of the predicted mask and the ground truth mask. Skeletonized masks reduce each segmented structure to its one-pixel-wide centerline, thereby preserving its overall topology and branching pattern while removing variations in thickness. Predictions will have a low loss if their skeletons overlap well with the skeletonized target mask and have a high loss if they do not. In a more mathematical sense, let y represent the ground-truth mask and $\hat{y}$ the predicted binary mask, from which the skeletonized masks S_y and $S_{\hat{y}}$ are derived. From these and our initial masks, we can calculate what the authors define as topological precision and sensitivity using the formulas below.

$$T_{\text{prec}}(S_{\hat{y}}, y) = \frac{|S_{\hat{y}} \cap y|}{|S_{\hat{y}}|}, \qquad T_{\text{sens}}(S_y, \hat{y}) = \frac{|S_y \cap \hat{y}|}{|S_y|}.$$

Finally, we take the harmonic mean of these values to get the ClDice score.

$$\text{clDice}(\hat{y}, y) = 2 \times \frac{T_{\text{prec}}(S_{\hat{y}}, y)\, T_{\text{sens}}(S_y, \hat{y})}{T_{\text{prec}}(S_{\hat{y}}, y) + T_{\text{sens}}(S_y, \hat{y})}.$$

Betti Components

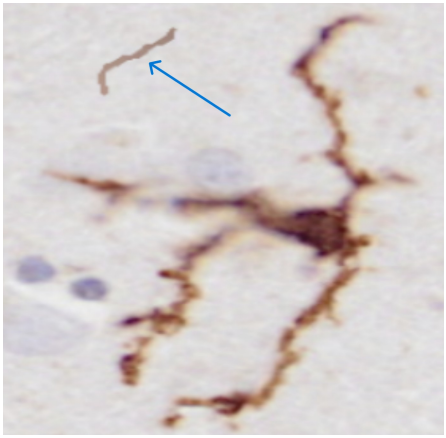
ClDice encourages the model to produce matching branches, but it does not enforce the correct number of connected components. For example, very small gaps in the skeleton may not significantly affect the ClDice or BCE but can have a large impact on downstream analyses. Thus, a final loss component was included that calculates the number of connected components in the target and predicted masks and penalizes discrepancies between them. This helps reduce fragmented predictions and encourages the segmented microglia to remain structurally connected.

5.3.3 Augmentations

Data augmentations are techniques frequently used in computer vision tasks to improve the robustness of models, reduce overfitting, and increase the effective diversity of the training data [36]. It involves expanding the training dataset by applying transformations to existing training samples that alter their appearance while preserving their validity as training examples. In this work, we applied several standard data augmentations, including resizing all images to 256 * 256 pixels, random horizontal flips, random vertical flips, random 90-degree rotations, color jitter, Gaussian blur and Gaussian noise.

In addition to these standard data augmentations, we have also implemented two custom transformations specific for microglia segmentation. In particular, these were designed to make the model more robust to the data issues associated with converting 3D microglia into 2D scans, as discussed in 4.1. Firstly, we have added synthetic cuts in the processes (branches) of the microglia by replacing a small portion of the process with a random section of background. The random background was selected rather than plain white to ensure that the model did not overfit to these new synthetic changes. Secondly, we added fragments of other microglia that are too far from the target cell to be included in its mask. Figure 9 shows an example of both the synthetic cuts in (a) and added fragments in (b).

(a) Blue arrow shows a synthetic cut example.



(b) Blue arrow shows a synthetic fragment example.

Figure 9: Examples of the two custom domain-specific augmentation strategies used to generate additional training examples. This reduces overfitting and allows the model to become more robust to issues in the scans caused by slicing a 3D cell into 2D.

5.3.4 Postprocessing

Although the U-Net model produces strong initial predictions, it frequently identifies multiple disconnected components for the same cell. This should not occur because a single microglial cell should be represented as one connected object. To address this issue, two competing post-processing techniques were developed to calculate the optimal path for connecting the different components.

First, an image-processing-based approach calculated the cost of connecting two pixels by assigning a lower cost to paths that passed through brown pixels. Second, a logits-based approach used the U-Net output logits as the cost map and then calculated the optimal path for connecting two components.

Once an optimal path was identified, the thickness of the connection was calculated dynamically based on the thickness of the two components. Figure 10 shows an example of this post-processing in action, and Section 7.5.2 shows the performance of the different post-processing methods in terms of how well they improve the predicted morphologies.

It is important to note that components that were considered too far from the main cell were not connected, as these were deemed to be fragments of other cells that were not fully captured in this scan. See Section 4.1 for more information on this.

(a) Example segmentation before postprocessing

(b) Example segmentation after postprocessing

Figure 10: Microglia segmentation example both before connecting the distinct components and after using our U-Net logits based approach.

5.4 Morphological Features

Based on the segmentation mask produced by our U-Net, we calculated 25 morphological features that quantify the shape of the microglia. These features are summarized in Table 1 and Appendix 11.1 gives a visual overview of how each feature works. The feature selection was inspired by those in [42] to capture a diverse range of differences between cells. Overall, they can be grouped into five categories: skeleton-based topology features, Sholl analysis features, soma features, whole-cell features and branch features.
Sholl features are a way to measure the complex branching structures frequently seen in microglia and other cells. They were first introduced by Sholl in [35] to analyze cat neurons by quantifying how branching complexity changes as a function of distance from the center of the cell. The intuition behind a microglial Sholl analysis is that concentric circles are drawn around the soma, and the number of times the cell skeleton intersects each circle is counted. From these intersection counts, several summary features can be derived including the minimum radius, peak radius, max radius and sum. See Table 1 for more information.

Table 1: Morphological features used to quantify microglial shape.

Feature	Description
skeleton_length	Total length of the skeletonized cell processes, measured as the number of skeleton pixels.
num_junctions	Number of distinct branch points in the skeleton. A junction is identified by being a skeleton location with at least 3 direct neighboring pixels.
num_components	Number of connected components in the cell mask. Typically this is only one as a cell should be fully connected.
num_end_nodes	Number of terminal skeleton points, ie. pixels with only one neighbor in the cell.
num_start_nodes	Number of points that connect the processes to the soma. These are identified by being skeleton points with only one neighbor and are touching the soma.
total_nodes	Total number of terminal and starting nodes.
end_to_start_ratio	Ratio between the number of end nodes and start nodes.
soma_area	Area of the soma mask.
soma_perimeter	Perimeter of the soma mask.
soma_circularity	Measure of how close the soma shape is to a perfect circle, computed as $4\pi A/P^2$, where A is the soma area and P is the soma perimeter.
cell_area	Area of the full cell mask.
cell_perimeter	Perimeter of the full cell mask.
cell_convex_hull_area	Area of the convex hull enclosing the full cell mask.
convex_hull_perimeter	Perimeter of the convex hull enclosing the full cell mask.
cell_solidity	Proportion of the convex hull occupied by the cell, reflecting how concave or ramified the shape is. Computed as $A_{\text{cell}}/A_{\text{hull}}$, where A_{cell} is the cell area and A_{hull} is the convex hull area.
cell_convexity	Ratio between the convex hull perimeter and the cell perimeter, reflecting boundary irregularity. Computed as $P_{\text{hull}}/P_{\text{cell}}$, where P_{hull} is the convex hull perimeter and P_{cell} is the cell perimeter.
cell_circularity	Measure of how close the full cell shape is to a perfect circle, defined as $4\pi A_{\text{cell}}/P_{\text{cell}}^2$, where A_{cell} is the cell area and P_{cell} is the cell perimeter.
cell_convex_circularity	Circularity of the convex hull of the full cell mask, defined as $4\pi A_{\text{hull}}/P_{\text{hull}}^2$, where A_{hull} is the convex hull area and P_{hull} is the convex hull perimeter.

Feature	Description
branch_area	Area of the branch region which represent the cell processes.
branch_perimeter	Perimeter of the branch region which represents the cell processes.
sholl_min_radius	Smallest radius from the soma centroid at which the skeleton intersects a Sholl ring.
sholl_peak_radius	Radius from the soma centroid at which the maximum number of Sholl intersections occur.
sholl_max_radius	Largest radius from the soma centroid at which the skeleton intersects a Sholl ring.
sholl_peak	Maximum number of skeleton intersections observed across all Sholl radii.
sholl_sum	Total number of skeleton intersections summed across all Sholl radii.

5.5 Dimensionality Reduction

Before clustering the microglia to identify morphologies, it is essential to first reduce the dimensionality. This is because many of the morphological features measure similar underlying ideas and thus are very correlated with each other. For example, branch skeleton length is highly correlated with branch area, as both quantify branchiness. Therefore, clustering directly on the morphological features would effectively count branchiness twice, since several features measure the same underlying property in different ways. See Appendix 11.1 for visual overviews of all morphological features.

Principle Component Analysis (PCA)
PCA is a linear dimensionality reduction technique commonly used in a variety of biological analyzes. The goal is to represent the input data using fewer variables while preserving as much variance as possible [17]. This is done by first scaling the input features followed by computing their covariance matrix. Next, by performing eigendecomposition, we can identify the principal components, which are the directions in feature space along which the projected data has maximum variance. Next, we needed to determine the number of principal components required to adequately represent the morphological features. If we selected too many, the dimensionality would not be reduced enough to address the correlation issue highlighted above, while too few PCs would insufficiently represent our morphological features. Therefore, the elbow method (Figure 11) was used to strike this balance, yielding four principal components that captured 80% of the variance in the input features.

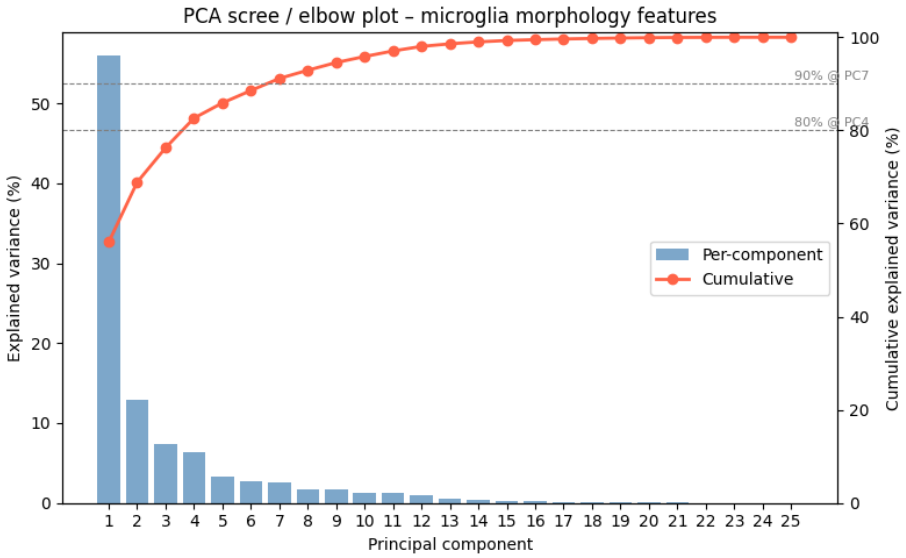


Figure 11: PCA scree plot showing how the cumulative explained variance of morphological features changes with increasing principal components. We see that only 4 principal components can already represent 80% of the variance present in the morphological features and so for downstream tasks only 4 PCs were used.

t-distributed Stochastic Neighbor Embedding (t-SNE)
While PCA is effective, it can only capture linear relationships between features, meaning it may perform suboptimal dimensionality reduction if nonlinear relationships are present. Therefore, t-SNE was also implemented as an alternative dimensionality reduction technique to assess whether it better captures the structure of the morphological features. While it shares the same goal as PCA, namely dimensionality reduction, t-SNE relies on fundamentally different mathematical principles to achieve this. In essence, t-SNE works by preserving the local distances between high-dimensional neighbors in the low-dimensional space. It identifies these neighbors by calculating the pairwise distances of all points in the original feature space and then converting these to probabilities. It then places points in a low dimensional space and iteratively updates their locations to ensure that points with a high similarity in the original feature space also have a high similarity in the target space. The number of dimensions used for this work was set to 4, so as to match the number of principal components identified above and thus make comparisons more robust.

5.6 Clustering

After the morphological feature extraction and subsequent dimensionality reduction, we used clustering to identify the different microglia morphologies present in our data. This is motivated by prior literature, which has found that microglia fall into a set of morphologies such as Homeostatic, Amoeboid, and Rod, among others [28]. As our scan data is unlabeled, clustering was used to discover these morphological groups without relying on predefined class labels. The specific hard clustering techniques compared in this work are KMeans, Agglomerative clustering and HDBScan. Next, soft clustering techniques were also used to identify morphologies, with these assigning each data point a proportion of membership to each cluster. For example, a data point may belong 55% to cluster 1, 20% to cluster 2, and 25% to cluster 3. Compared to in hard clustering where it would only be assigned to cluster 1. This is particularly appropriate for analyzing microglia morphologies, as they are naturally continuous and overlapping [4]. Furthermore, the two specific soft clustering techniques implemented were FCMeans and GMM. See section 2.3 for more information on how these techniques work and why they were selected. To our knowledge, this is the first microglia morphology analysis to model cell states as soft clusters, representing a novel contribution of this work.

5.7 Statistical Testing

The final stage of our pipeline is a set of statistical tests that assess if there are differences in microglia morphology between the diagnosis groups. These included a mixed effect model for comparing individual morphological features, a PERMANOVA test for examining multiple features combined and finally, a Dirichlet analysis for assessing cluster proportions.

5.7.1 Mixed Effect Modeling

To determine if there are differences in individual morphological features between the diagnosis groups, we conducted a mixed-effects analysis. We first calculated the mean value of each morphological feature per patient and then analyzed whether these means differed across diagnosis groups. However, there are other exogenous variables like sex that can have an impact on our results, so a general test is unsuitable. To account for this, we fit a linear model that estimates the effects of the diagnosis groups and the other covariates, thereby allowing us to assess whether the diagnosis groups truly differ. In mathematical terms, we predict a given morphological feature y_i using the equation below, where β_0 is the intercept, β_1 represents the effect of diagnosis group, β_2 represents the effect of sex, and ε_i is the residual error term. A large value of β_1 indicates a stronger effect of diagnosis group on the response variable which we can use to assess if there is a significant difference between the diagnosis groups.

$$y_i = \beta_0 + \beta_1 \,\text{Diagnosis}_i + \beta_2 \,\text{Sex}_i + \varepsilon_i$$

Since we are performing this test across multiple morphological features, we run the risk of p value hacking. Therefore, we also calculate the False Discovery Rate (FDR) and quantify this with a q value. However, since many of the morphological features are correlated with one another, this can artificially make the results appear less significant due to testing for the same underlying property in multiple tests. To overcome this, we have chosen to only perform our mixed effect modeling test on a subset of the features with these being uncorrelated with one another. Specifically we tested on the following features: cell_convex_circularity, num_components, sholl_min_radius, sholl_sum and soma_perimeter. To select these features, we used correlation as an inverse distance measure, clustered the features based on this distance, and then selected the feature closest to the centroid of each cluster.

5.7.2 PERMANOVA

We also implemented a statistical test to explore if there are differences between diagnosis groups based on a combination of morphological features. Each individual was summarized by the mean value of their morphological features, as described in 5.7.1. Next, using the same logic as with our mixed-effect modeling test, we only test on a subset of the morphological features with these being cell_convex_circularity, num_components, sholl_min_radius, sholl_sum and soma_perimeter. The PERMANOVA test itself models each individual as a vector of morphological features and then compares whether these vectors are grouped according to diagnosis group. It does so by first building a Euclidean pairwise distance matrix and then calculating whether the within-group distances are smaller than the between-group distances. Finally, to assign significance to this difference, it performs permutations in which different individuals have their diagnosis labels swapped. After each permutation, the within-group and between-group distances are recalculated. If they differ greatly from the initial calculation, then the diagnosis group must be an important factor, resulting in a low p-value.

5.7.3 Dirichlet Analysis

The research question this analysis aims to answer is the following: Are there differences in cluster proportions and their variance between diagnosis groups? In order to explore this, we must first calculate per patient what proportion of their cells came from each cluster, meaning each individual is represented by a single compositional vector regardless of how many cells were present in their scan. Next, since the cluster proportions are all greater than 0 and add up to 1, we can model them together as Dirichlet distributions. This allows for a richer analysis than comparing each cluster proportion individually, since if one cluster proportion increases, the others must decrease. This relationship would be lost if we compared only the mean cluster proportions independently. Specifically, our Dirichlet distributions work by estimating one parameter per cluster, resulting in a vector $\alpha = (\alpha_1, \ldots, \alpha_K)$. Larger values of α_1 indicates that cluster 1 has a

higher proportion than the other clusters. The sum $\sum_i \alpha_i$ represents the inverse of the variance as a larger sum means a tighter distribution. Figure 12 shows an example of how varying the α values impacts the distribution of the data. The goal of this analysis is to model three Dirichlet distributions, one for each diagnosis group, and to assess whether the groups differ by testing for significant differences in their α parameters. This is done both individually and in their overall sum, which reflects variance.

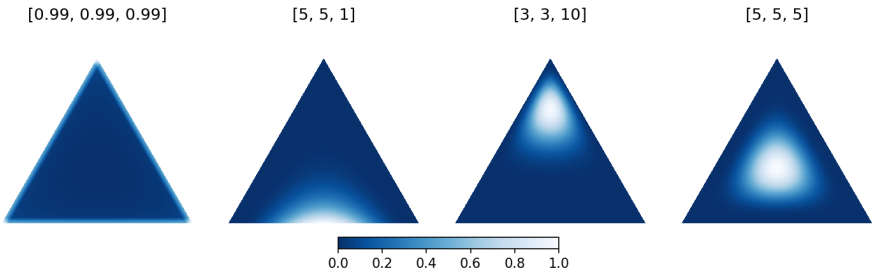


Figure 12: Four Dirichlet distributions with different α values showing how each parameter effects the overall distribution. Figure originally introduced in [9]

6 Experiments

6.1 Cell Detection

One of the main limitations present in previous microglia detection work is that they use their training data very inefficiently, meaning they need a very large training set to achieve strong performance. This work aims to address this limitation by achieving strong results with a much smaller dataset through the use of both active learning strategies as well by applying transfer learning from rat microglia detection. Thus, we have constructed several experiments to quantify the performance impact of these techniques to determine which are most suitable for microglia detection.

Firstly, as a baseline, a YOLOv8 model was initialized from standard pretrained weights and then trained on only human microglia data using 5 fold cross validation. For each train-test-val split, 70% of data was used for training, 10% for validation and finally 20% used as the testing set. By doing these multiple splits, we ensure that our model truly performs well on unseen data rather than relying on a favorable split. Furthermore, the validation set was not used to tune any hyperparameters; instead, it was used for patience-based early stopping to prevent the model from overfitting on the training data.

Secondly, we assessed the performance of first pretraining on rat microglia before finetuning on human data. For this, we used the exact same training and evaluation setup as described above. This allows us to assess the performance difference achieved by the inclusion of rat data and determine if inter species

transfer learning can lead to better microglia detection performance.

Thirdly, a learning curve experiment was conducted where a YOLO model was trained on different amounts of training data while being evaluated on the same test set. This allowed us to quantify the effect of training data size on our performance. This was critical, as labeling additional data is very time-consuming and only logical if it results in a performance improvement — a decision that can be informed by the trajectory of this learning curve.

Finally, we assessed the performance of the different active learning approaches described in 5.2 using two different methods. First, we used all active learning datasets as additional training data to assess which active learning method resulted in the greatest performance gains. In addition to this, we assessed the model's performance when trained normally but evaluated on the morphologically diverse dataset. This can be seen as an indicator for the worst case performance of our model on unseen data, since the cells are very morphologically different than what was present in the training data.

To evaluate the model's performance, we used precision, recall, mAP@50 and mAP@50-95. Firstly, precision and recall give us a rough indication for how many of the microglia are being identified or missed. However, they do not take into account what proportion of the predicted bounding box overlaps with the ground truth. They are defined below with TP, FP and FN denoting the numbers of true positives, false positives, and false negatives, respectively. Next, $AP^{(t)}$ is the average precision at IoU threshold t, computed as the area under the precision–recall curve. This gives us a greater insight into how many of the cells are actually being identified with a good bounding box. For example, with mAP@50 we are able to assess what proportion of the cells that our model identifies overlap at least 50% with the target. Finally, mAP@50-95 is the mean average precision aggregated across classes and IoU thresholds from 0.50 until 0.95 in increments of 0.05. This provides a stricter measure of localization quality than mAP@50 and gives an indication of how many predictions are truly accurate, rather than simply achieving an overlap of 0.5 with the target.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$AP^{(t)} = \int_0^1 p_c^{(t)}(r)\, dr \qquad \text{mAP@50} = \frac{1}{C} \sum_{c=1}^{C} AP_c^{(0.50)}$$

$$\text{mAP@50-95} = \frac{1}{10C} \sum_{c=1}^{C} \sum_{t \in \{0.50, 0.55, \ldots, 0.95\}} AP_c^{(t)}$$

6.2 Segmentation

Previous microglia segmentation models share two key limitations. First, they employ very few custom microglia-specific techniques to improve performance. Furthermore, their evaluation is lackluster, meaning it is unclear how well their models preserve biologically meaningful properties, such as the morphological features and the complex branching structure of microglia. Thus, we have implemented our own custom U-Net model and have designed several experiments to assess its performance compared to foundational segmentation models as well as the impact of our custom microglia-specific techniques. All experiments are conducted using a robust evaluation framework, including a custom morphology metric that allows us to quantify our model's ability to preserve morphological features.

Firstly, a custom U-Net was implemented using a fully supervised setup and 5 fold cross-validation, where 70% of the data was used for training, 20% for testing, and 10% for validation. This model's performance was then compared quantitatively to the Segment Anything Model (SAM) with box based prompting [14]. This allows us to measure the performance gained from training a custom segmentation model compared to one trained for a variety of tasks. Following this, we performed one experiment for every custom microglia-specific step in our segmentation to assess if it improves performance. To begin with, an augmentation ablation study was conducted to evaluate whether the custom disconnection and synthetic fragments described in Section 5.3.3 improved robustness. Next, we assessed the impact of using different combinations of loss components to determine how each of them contributes to the model's final performance. Finally, we quantified the impact of postprocessing the initial segmentation by connecting components through both a logits-based method as well as one based on an image processing technique. See Section 5.3.4 for more information.

The evaluation for our segmentation experiments was done using three complimentary techniques. Firstly, we assessed it using Dice and IoU, which are common segmentation metrics used for a variety of different computer vision tasks. They are defined using the formulas below where y represents the ground truth mask and $\hat{y}$ is the predicted mask.

$$\text{Dice}(\hat{y}, y) = \frac{2|\hat{y} \cap y|}{|\hat{y}| + |y|} \qquad \text{IoU}(\hat{y}, y) = \frac{|\hat{y} \cap y|}{|\hat{y} \cup y|}$$

Secondly, these traditional metrics were calculated both at the whole cell level as well as broken down by soma vs processes. This allows us to assess how well the segmentation performs across different sections of the cells and is a step missing from the vast majority of previous research done in this topic.

Finally, the models were evaluated on how well they predicted each of the morphological features. This was done by calculating the morphological features for both the predicted and target masks, then computing the absolute percent-

age error for each feature. This evaluation is critical because it provides granular insights into the strengths and shortcomings of the models. Additionally, it gives a measure of confidence in the downstream analyses, since those results become meaningless if the morphological features are not predicted correctly. However, with there being 25 morphological features, it is difficult to assess which model version has the best overall morphological performance. One solution could be to calculate a custom morphology score as the mean error across all of the features. Unfortunately, this is unsuitable as some features are much more important than others, eg. branch length is very important. To address this limitation, we instead opted for a weighted average of the feature errors, with the weights representing the feature importance for our downstream clustering. Since KMeans on t-SNE is not interpretable, we instead trained a random forest to predict the cluster label based only on its morphological features and then used SHAP to determine feature importance.

SHAP is an explainability technique based on an additive model, where each prediction is represented as a baseline value plus the sum of the feature contributions called SHAP values. Each SHAP value measures how much a given feature shifts the prediction away from its baseline value. They are calculated by training models on all possible combinations of features and measuring their marginal contribution [22]. Although SHAP values are only calculated for each individual data point, one can estimate the global feature importance by calculating the sum of the absolute SHAP values across all data points. Figure 13 shows these sums normalized to add up to 1, enabling us to determine the importance of the morphological features for predicting cluster labels.

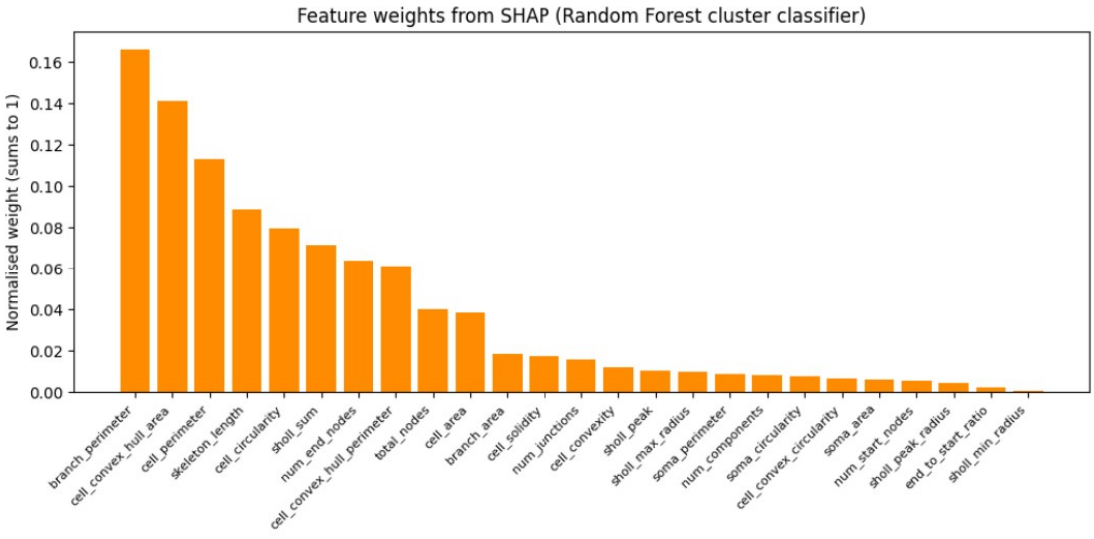


Figure 13: Shows the feature importance for clustering by taking the sum of absolute SHAP values across all data points.

6.3 Dimensionality Reduction

As discussed in Section 5.5, dimensionality reduction is a necessary step before clustering to avoid giving greater importance to correlated features. However, there are many possible dimensionality reduction techniques that could be appropriate for our task, with the vast majority of previous papers only implementing PCA [42]. We aim to determine whether the non-linear dimensionality reduction technique t-SNE is more suitable than PCA for capturing morphological variance in a reduced dimensional space. To do so, the methods were assessed using two neighborhood-preservation metrics: trustworthiness and continuity. Trustworthiness is similar to precision in that it measures the proportion of local neighbors in the low-dimensional space that were also true neighbors in the high-dimensional space. Contrastingly, continuity is similar to recall in that it quantifies what proportion of the neighbors in the high dimensional space were preserved as neighbors in the low dimensional space.

Each term is defined more precisely using the formulas below:

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_k(i)} \big(r(i,j) - k\big)$$

$$C(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in V_k(i)} \big(\hat{r}(i,j) - k\big)$$

- n is the number of data points.
- k is the number of neighbors considered.
- $U_k(i)$ is the set of points that are neighbors of point i in the low-dimensional space but not in the original space.
- $V_k(i)$ is the set of points that are neighbors of point i in the original space but not in the low-dimensional space.
- $r(i,j)$ is the rank of point j among the neighbors of i in the original high-dimensional space.
- $\hat{r}(i,j)$ is the rank of point j among the neighbors of i in the low-dimensional space.

6.4 Clustering

As discussed in Section 2.3, there are many possible hard clustering methods that could be suitable for capturing microglia morphologies. However, all previous literature has simply selected a single clustering method for their analysis without determining first if it was truly optimal for the task [42], [39]. We aim to overcome this shortcoming by performing two experiments to determine which clustering method is most suitable for identifying microglia morphologies. Firstly, we determined the optimal number of clusters k by assessing which k

value led to the highest silhouette score. Secondly, we compared the different clustering methods to one another to determine which was best for the task. A silhouette score ranges from -1 to 1 and encourages points to be similar to those within its cluster compared to other clusters. Mathematically, it is defined using the formula below where $a(i)$ is the average distance to points within its cluster, and $b(i)$ is the average distance to points in the nearest other cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

In addition to this unsupervised validation technique, we also assessed how well the different clusters agreed with manual annotations from an expert. To do so, 97 cells were manually assigned to one of the following microglial morphologies: Homeostatic, Reactive, Amoeboid, Hyper-ramified, and Rod (see Section 2.1 for more information on what these represent). Next, Normalized Mutual Information (NMI) was used to assess the agreement between the clusters and the manual annotations, with a contingency table also enabling us to better understand where the differences occurred.

Along side this, we also conducted an experiment to assess which soft clustering method was most appropriate for grouping microglia based on morphology. The two soft clustering techniques that were compared are Fuzzy C-Means and GMM. To evaluate which method is most suitable, we calculated the mean entropy of their cluster assignments. If a model is consistently assigning very high cluster proportions, then it will have a low entropy. Mathematically, it is defined using the function below with N being the number of cells, K the number of clusters and p_{ik} the probability of sample i belonging to cluster k. In most machine learning tasks, a low entropy is preferred. However, in this work, we have instead chosen to encourage soft clustering methods that have a high entropy (within reason). This is for two reasons. Firstly, as seen in Section 7.4, microglia lie on a morphological spectrum rather than in say in distinct states with very few transition microglia between these states. Since these concrete cell states are not present, it is best to reward methods that spread its predictions more evenly across the clusters when appropriate. Secondly, in Section 7.5, we analyze the differences between diagnosis groups for both soft and hard clusters, meaning that if the clusters have too low of an entropy, they would essentially be the same as the hard clusters.

$$\bar{H} = \frac{1}{N} \sum_{i=1}^{N} \left(- \sum_{k=1}^{K} p_{ik} \log_2 p_{ik} \right) \tag{1}$$

6.5 Statistical Analysis

As discussed in Section 5.7, our individuals were distributed across three diagnosis groups: 100+_with_pathology, 100+_without_pathology and AD. All diagnosis groups were compared to one another in a pairwise fashion to determine if there were differences between the groups in cluster proportions, individual morphological features and combinations of morphological features. See Section 5.7 for for the full details on the statistical tests applied to detect these differences. Each pairwise comparison allows us to both understand how AD affects the brain and the longevity present in the centenarians. Below is a summary of the motivations and insights available from each pairwise comparison.

AD vs 100+_with_pathology
Both of these groups had the AD pathology, yet the centenarians managed to live over 100 years, while those in the AD group died significantly earlier. By comparing the microglial morphological profiles between these groups, we aim to identify whether differences in microglial cell states may help explain this disparity in outcomes. By understanding and quantifying these differences, future research may be able to develop treatments that shift living AD patient's microglial profiles toward those observed in cognitively resilient centenarians. This could ultimately slow disease progression and improve patient outcomes.

AD vs 100+_without_pathology
Comparing these two groups allows us to understand how the microglial morphology differs between AD patients and healthy individuals who lived beyond 100 years. While causation cannot be directly established, this comparison still enables us to identify which morphological profiles are associated with AD and, conversely, which are characteristic of healthy longevity. By understanding these differences, future research could investigate treatments to make the general population's microglia more like the healthy centenarians, enabling us to live longer and delay the onset of AD.

100+_with_pathology vs 100+_without_pathology

Comparing these two group reveals the diversity within the centenarian cohort. Although all of these individuals achieved an advanced age, they did so under very different circumstances as some of them had the AD pathology while others were healthy. This raises the question of whether the presence of AD can be linked to specific microglia morphological profiles? If this is the case, then the profile associated with 100+_with_pathology may be specifically well suited for longevity while combating AD. Contrastingly, if the two centenarian groups have very similar morphological profiles, this could indicate that there is a microglial 'silver bullet' linked to longevity. This profile would be truly unique and one that is characteristic of an extraordinary lifespan regardless of whether they develop AD.

7 Results

7.1 Cell Detection

As discussed in Section 6.1, the goal of our detection experiments was to answer four main questions:

- How does pretraining on rat cells improve performance on human microglia detection?
- How does our YOLO performance change with different amounts of training data?
- Which active learning technique is most suitable for microglia detection?
- How does the performance of our YOLO model change when evaluated on a test set that is selected to be morphologically different from its training data?

To assess the impact of pretraining on rat data for human microglia detection, we trained two different YOLO models. First, one trained on only human data and second, using rat pretraining followed by human fine-tuning. Table 2 shows a comparison of their performance on the unseen test sets. We can see that both YOLO models perform very well for microglia detection with mAP@50s of over 0.87. For comparison, StainAI's pipeline achieved a mAP@50 of only 0.796 on rat data. While the results are not directly comparable due to differences in dataset and experimental setup, this still suggests that our detection component is performing well. Next, for both models, the recall is substantially lower than the precision, indicating that false negatives are a weak point of our current approach. Furthermore, the significant drop between mAP@50 and mAP@50-95 suggests that while our model correctly detects most cells, the predicted bounding boxes are not always tightly aligned with the targets.

When comparing the two models, we see that pretraining on the rat data improves performance significantly with mAP@50, mAP@50-95 and Precision all improving by between 1.5% and 2%. Additionally, we found that this performance improvement was present across every train-test split, indicating that this difference is robust. Interestingly, we do not see a significant change in the Recall. One possible explanation for this is that rat microglia do not add substantially new morphological variation beyond what is already present in the human training data.

Model	mAP@50	mAP@50-95	Precision	Recall
YOLO only human data	0.870	0.508	0.817	0.798
YOLO rat pretraining	0.886	0.528	0.832	0.803

Table 2: Detection performance of two competing YOLO models. 'YOLO only human data' was trained for 100 epochs then evaluated on a human test set while 'YOLO rat pretraining' first had 100 epochs of training on rat microglia then 100 epochs of fine tuning on the human training data before being evaluated on the human test set.

To assess how detection performance scaled with the amount of available training data, we conducted a learning curve experiment for the YOLO model. Note that this was done on the original YOLO model, meaning without first pretraining on rat data so that we can assess each experiment's individual impact. Figure 14 shows the performance of the YOLO model when trained on different proportions of the training data and evaluated a single test split. Multiple different splits were taken for the training data proportion, eg. different splits of 10% were used as training data and then evaluated on the same test set. This was done to account for variability caused by the particular choice of training split. The figure shows both the mean performance as well as the minimum and maximum split performances. We see that performance improves quickly and even when using just 50% of the data, our YOLO model can accurately detect microglia with a mAP@50 of over 0.8. However, when we increase the size of the training dataset further, we see that the performance is still improving but very slowly. This indicates that our model's performance is reaching the point of diminishing returns with respect to additional random cell annotations. Thus, showing the need for our active learning strategies, since based on this current trajectory, we would need several thousand more randomly selected annotations to improve performance significantly.

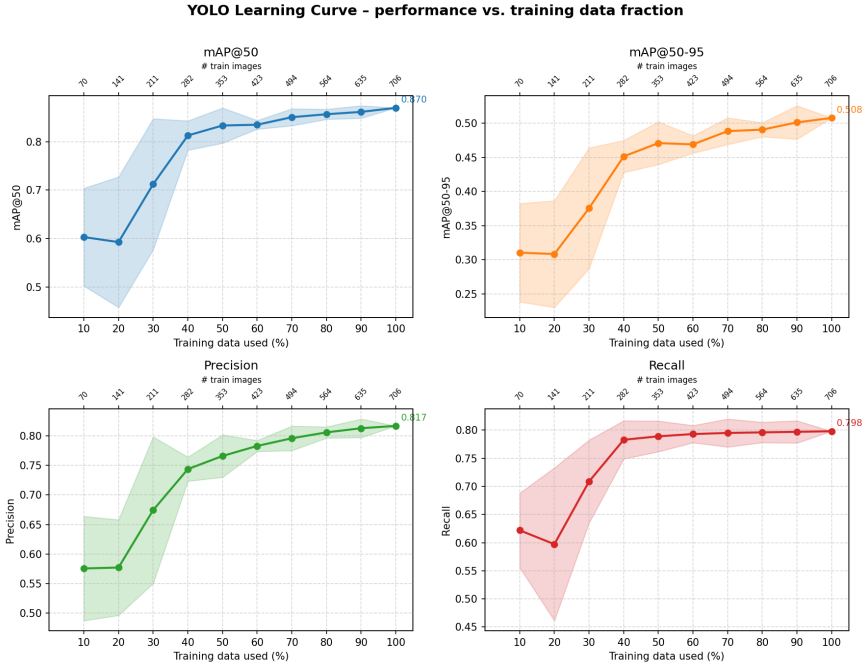


Figure 14: YOLO learning curve evaluated against the same test set with increasing proportions of training data. Five folds were used for training set selection to account for split variability; the graph shows the mean performance per split as well as the minimum and maximum. For the final 100% split, results across all five test splits were used, as these were deemed more representative of the model's performance on unseen data.

As introduced in Section 6.1, we conducted an experiment to determine which active learning strategy is best suited for microglia detection. Table 3 shows a comparison of the methods and their performance with the additional annotations added to the training split only. Unsurprisingly, providing additional randomly selected cells did not significantly improve model performance, suggesting that the model has sufficiently learned to identify general microglia. However, we did see significant performance gains when incorporating additional training that focuses on edge confidence and morphologically diverse cells. See Section 5.2 for more information on how these techniques work. Specifically, we found the greatest performance improvement was achieved by selecting the most morphologically diverse cells with this increasing mAP@50 by 0.7% and Precision by 1.7%. To the best of our knowledge, this morphology based active learning strategy represents a novel contribution to the field of microglia detection. Moreover, these results suggest that future microglia detection studies would benefit significantly from incorporating this approach, with potentially even greater gains achievable by applying it earlier in the annotation process, i.e., with fewer initial labeled samples.

Active learning data selection method	mAP@50	mAP@50-95	Precision	Recall
Original Dataset	0.870	0.508	0.817	0.798
Random	0.871	0.512	0.820	0.801
Edge Confidence	0.875	0.521	0.827	0.802
Morphologically diverse	**0.877**	0.517	**0.834**	0.803

Table 3: YOLO detection performance when incorporating additional training data obtained from different active learning methods. Edge Confidence were cells that the original YOLO model had a confidence prediction of between 0.45 and 0.55. The Morpholically diverse cells were selected to be as morphologically different than the original dataset.

Next, we performed an experiment to assess how well our detection model would perform on the morphologically diverse dataset as the test set. This evaluation allows us to assess how our model performs on unseen data in a worst-case scenario ie. one where the testing data is very different from the annotated training data. Table 4 shows the results of this analysis. Overall, we can see that while performance is slightly lower on the morphologically diverse dataset, it remains strong, with both box quality metrics decreasing by less than 0.5%. Furthermore, while Recall fell by 0.9%, Precision remained strong at 0.819. Such a minor decrease in performance indicates that our model has not overfit to its training data and is able to accurately detect cells that are morphologically dissimilar to those encountered during training.

Active learning data selection method	mAP@50	mAP@50-95	Precision	Recall
Original Dataset	0.870	0.508	0.817	0.798
Morphologically diverse	0.865	0.505	0.819	0.779

Table 4: YOLO detection performance when evaluated on the Morphologically Diverse dataset as the test set. The morphologically diverse dataset contains cells that are as morphologically different as possible than the original data set.

In summary, we found that human microglia detection performance can be improved by approximately 2% by first pretraining on publicly available rat microglia. Furthermore, through our learning curve analysis, we observed our model had reached the point of diminishing returns on random data, highlighting the need for greater data efficiency going forward. We addressed this in our next experiment by comparing multiple active learning strategies. The most suitable strategy was found to be our custom method for selecting morphologically diverse cells, with it outperforming both random data selection and the traditional edge confidence based approach. Finally, when evaluating our model on cells that are morphologically different than those in our training set, we saw strong performance, indicating our YOLO performs well on truly new data.

7.2 Segmentation

As discussed in Section 6.2, our segmentation experiments aimed to answer the following main questions:

- How does a custom U-Net compare to the foundational segmentation model SAM?
- How well does our U-Net predict individual morphological features and morphology as a whole?
- Do custom microglia-specific data augmentations improve segmentation performance beyond traditional augmentation strategies?
- How do different loss components influence segmentation performance?
- Is a U-Net logits-based post-processing method more suitable than an image processing approach for microglia segmentation?

To assess whether our custom topology-aware U-Net had truly learned microglia specific cellular characteristics, we compared its performance to the foundational model SAM [14]. SAM has not been trained on microglia images but has been shown to perform well on a variety of medical tasks and therefor can be seen as a baseline for what can be achieved without a custom model [24]. Table 5 shows the segmentation performance of SAM and our final U-Net model with optimal augmentations, loss components and postprocessing selected. On the whole, we see that our model is able to produce strong masks, with it outperforming SAM on all measures. Furthermore, with a morphology score of over 90%, our model is able to accurately predict the most critical morphological features needed for downstream analyses. However, one weakness of our U-Net is its imbalanced performance, as it predicts the soma much more accurately than the branches. This is logical, as the soma is typically more circular and has clearer boundaries, while the branches are often thin and can have complex structures. Quantitatively, we see that the branch IoU and Dice are both significantly lower compared to when calculated across the entire cell. Furthermore, the recall is also lower compared to the performance on the soma, indicating our model is more likely to be missing a branches then portions of the soma.

Model	Dice	IoU	Morphology	Soma Recall	Branch Dice	Branch IoU	Branch Precision	Branch Recall
Custom U-Net	0.873	0.775	0.905	0.910	0.794	0.665	0.792	0.809
SAM	0.477	0.327	0.521	0.642	0.261	0.228	0.48	0.117

Table 5: Segmentation performance comparison between the custom U-Net and SAM with 'Morphology' here representing our custom morphology metric described in Section 6.2. Traditional metrics like Precision, Recall and IoU are reported on both the whole cell level as well as split across branches and soma, allowing us to observe the imbalanced performance of both the custom U-Net and SAM.

While accurate pixel segmentation is important, it does not by itself guarantee that the morphological features required for downstream analysis are predicted reliably. To address this, Table 6 shows our U-Net's ability to predict individual morphological features. The weight here represents the morphological feature importance, see Section 6.2 for more information on how these were calculated. For the most important and general features, we see that our model is very accurate with an error generally less than 10%. However, our U-Net sometimes struggles to predict fine-grained features related to the branching structure, particularly Sholl features and the counts of specific nodes and junctions. Taken together, the U-Net demonstrates strong predictive performance with respect to morphological features, which provides additional confidence in the validity of our downstream analyses, as these are heavily dependent on segmentation quality.

Feature	**Weight (%)**	**Performance (%)**	**Error**	**Contribution (%)**
branch_perimeter	16.62	92.76	0.0724	15.42
cell_convex_hull_area	14.15	89.79	0.1021	12.70
cell_perimeter	11.32	94.15	0.0585	10.66
skeleton_length	8.87	93.07	0.0693	8.25
cell_circularity	7.95	90.34	0.0966	7.18
sholl_sum	7.16	93.07	0.0693	6.66
num_end_nodes	6.40	83.49	0.1651	5.35
cell_convex_hull_perimeter	6.14	96.23	0.0377	5.91
total_nodes	4.03	82.93	0.1707	3.34
cell_area	3.85	94.40	0.0560	3.63
branch_area	1.85	87.87	0.1213	1.62
cell_solidity	1.76	91.48	0.0852	1.61
num_junctions	1.60	70.80	0.2920	1.13
cell_convexity	1.21	94.11	0.0589	1.14
sholl_peak	1.05	88.43	0.1157	0.93
sholl_max_radius	0.97	96.20	0.0380	0.93
soma_perimeter	0.88	85.00	0.1500	0.74
num_components	0.82	49.14	0.5086	0.40
soma_circularity	0.77	89.84	0.1016	0.69
cell_convex_circularity	0.67	96.33	0.0367	0.65
soma_area	0.61	75.86	0.2414	0.47
num_start_nodes	0.54	26.12	0.7388	0.14
sholl_peak_radius	0.44	79.28	0.2072	0.35
end_to_start_ratio	0.26	26.26	0.7374	0.07
sholl_min_radius	0.08	71.34	0.2866	0.06
TOTAL	100.00	–		90.05

Table 6: U-Net performance broken down by morphological feature. The weight here represents the feature importance for clustering outlined in Section 6.2 and the performance represents the model's accuracy in predicting a given morphological feature. The weight multiplied by the performance tells us how much this morphological feature contributes to the overall morphology score of 0.905.

Next, to assess wether custom augmentations would improve segmentation performance, we compared a U-Net trained with no augmentations to different combinations of both basic and custom augmentations. Table 7 shows the results for the most important combinations. Logically, the version trained without augmentations performs the worst, while even basic augmentations such as flips, rotations, crops, and noise improve the results significantly. Furthermore, both synthetic cuts and fragments improve morphology performance, but did not significantly affect Dice and IoU. This is logical, as these augmentations mostly affect the model's branch predictions, which are often thin and small structures whose changes greatly impact morphology. However, since they do not contain many pixels, they do not significantly change the IoU and Dice. Overall, these results are very promising and show that microglia segmentation is greatly improved by custom augmentations. Furthermore, this experiment highlighted the need for our custom morphology score, as both IoU and Dice would not have meaningfully identified this finding.

Augmentation Setting	**Dice**	**IoU**	**Morphology Score**
No augmentations	0.842	0.739	0.831
Basic augmentations	0.865	0.765	0.855
Basic + synthetic component	0.865	0.771	0.891
Basic + synthetic cut	0.868	0.772	0.895
All augmentations	0.873	0.775	0.905

Table 7: Comparison of segmentation and morphology performance with different augmentation settings. Basic augmentations implemented were flips, rotations, crops, gaussian noise and gaussian blur.

To determine whether topology-aware loss components improve segmentation quality, we compared multiple versions of our U-Net trained on different combinations of loss components. Table 8 shows the results of this analysis. As a whole, we see that BCE is needed for the model to achieve a strong Dice and IoU, since it penalizes predictions that are not aligned at a pixel level with the target mask. However, ClDice and Betti do improve the morphology score by approximately 2% when used as loss components compared to training only using BCE. This can mainly be attributed to ClDice's emphasis on preserving thin branching structures and Betti's encouragement of accurate topological features. Overall, these results are promising and indicate that custom loss components can play a significant role in accurately predicting morphological features in microglia.

Loss Component Setting	Dice	IoU	Morphology Score
BCE	0.8727	0.7786	0.8812
ClDice	0.6154	0.4457	0.7332
Betti	0.2100	0.1189	0.5427
BCE + ClDice	0.8675	0.7704	0.9016
BCE + ClDice + Betti	0.873	0.7726	0.9052

Table 8: Comparison of segmentation and morphology performance for different loss component settings. In BCE + clDice, BCE and clDice were assigned weights of 0.8 and 0.2 respectively, while in BCE + clDice + Betti, the loss component weights were 0.75, 0.2, and 0.05 respectively.

In our final experiment, we compared whether connecting the fragmented segmentation components into one object improved performance and if so, which technique was most suitable for the task. Table 9 shows the performance of the U-Net with and without the postprocessing methods described in Section 5.3.4. Interestingly, we see that both postprocessing methods improve the morphology score greatly. This is intuitive, as many morphological features are strongly influenced by which components are connected, e.g., the number of nodes, branching points, and branch lengths. However, both methods decreased the IoU and Dice scores by approximately 2%, indicating that, in some cases, components may have been connected to the main cell body through incorrect paths. Finally, the logits based method improved the segmentation score slightly more than the image processing approach. This indicates that our U-Net model's predictions are capturing useful structural information that is missed by the image processing approach, allowing it to recover more morphological details.

Postprocessing Method	Dice	IoU	Morphology Score
No postprocessing	0.873	0.775	0.810
Logits based	0.852	0.747	**0.905**
Image processing based	0.857	0.754	0.899

Table 9: Comparison of segmentation and morphology performance for different postprocessing methods. More information on how the custom morphology score is calculated can be found in Section 6.2

Overall, the main findings of our segmentation experiments are the following. Our custom U-Net vastly outperforms the foundational model SAM by as much as 40% on many metrics. However, it is somewhat biased with it performing better on segmenting the somas of the microglia than their complex branches. In addition, while its ability to predict individual morphological features does vary, as a whole it predicts them well, with a morphology score of over 90%. Furthermore, we saw significant performance improvements by incorporating custom data augmentations as well as topology aware loss components into our model. These findings highlight the importance of these custom microglia specific components which are novel contributions of this work. Finally, we found

that while both post-processing methods significantly improve morphological performance, the logits-based approach appears to be marginally more suitable, as it increases the morphology score by an additional 0.5% compared to the image processing approach.

7.3 Dimensionality Reduction

As discussed in Section 6.3, we performed an experiment to determine which dimensionality reduction technique was most suitable for capturing the morphological features in only 4 dimensions. Table 10 shows the performance of the competing methods t-SNE and PCA with the F-1 Score here being the harmonic mean of continuity and trustworthiness. Although continuity is slightly higher when using PCA, its trustworthiness is significantly lower than that of t-SNE, which leads to the large difference in their F1 scores. Overall, this indicates that t-SNE may be more suitable than PCA for capturing morphological features in microglia, even though PCA has been used in almost all previous literature [42], [39]. This is likely because the morphological features exhibit non-linear relationships with one another that t-SNE is better able to model.

Dimensionality Reduction	**Continuity**	**Trustworthiness**	**F1**
PCA	0.979	0.907	0.942
t-SNE	0.963	0.985	0.974

Table 10: Performance of t-SNE and PCA reducing the 25 morphological features down to only 4. The F1 score is calculated by taking the harmonic mean of continuity and trustworhiness.

7.4 Clustering

The goal of our clustering experiments was to determine which hard and soft clustering methods were most appropriate for capturing the complex microglia morphologies.

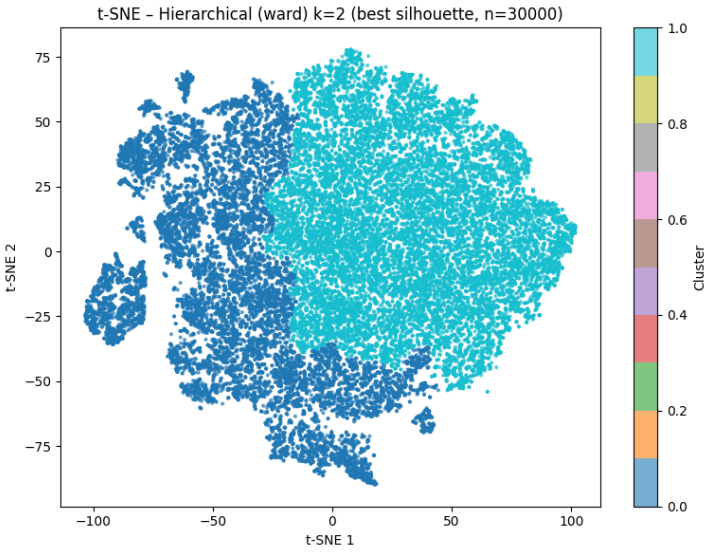
7.4.1 Hard Clustering

Before comparing the performance of the hard clustering methods against one another, we must first determine the optimal number of clusters for each method. For example, Table 11 shows the performance of K-Means with different numbers of clusters. We observed that optimal number of clusters was 4, with a Silhouette score of 0.3881. However, this finding was not consistent across all clustering methods with the HBDScan and Agglomerative clustering both identifying 2 clusters that broadly represent activated and non activated morphologies. This variability across clustering methods indicates that while a discrepancy between activated and resting microglia is robust, the other subdivisions may be finer and thus more reliant on specific clustering techniques to detect.

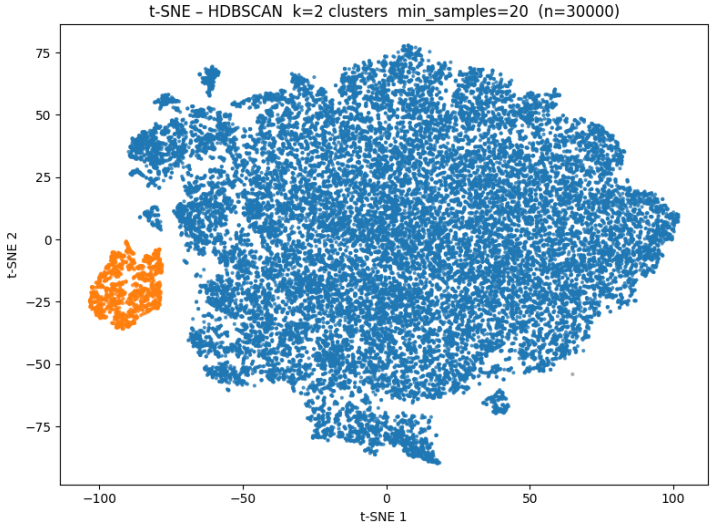
Number of clusters (K)	Silhouette score
2	0.341
3	0.357
4	**0.388**
5	0.321
6	0.305
7	0.301

Table 11: Silhouette scores for different numbers of clusters.

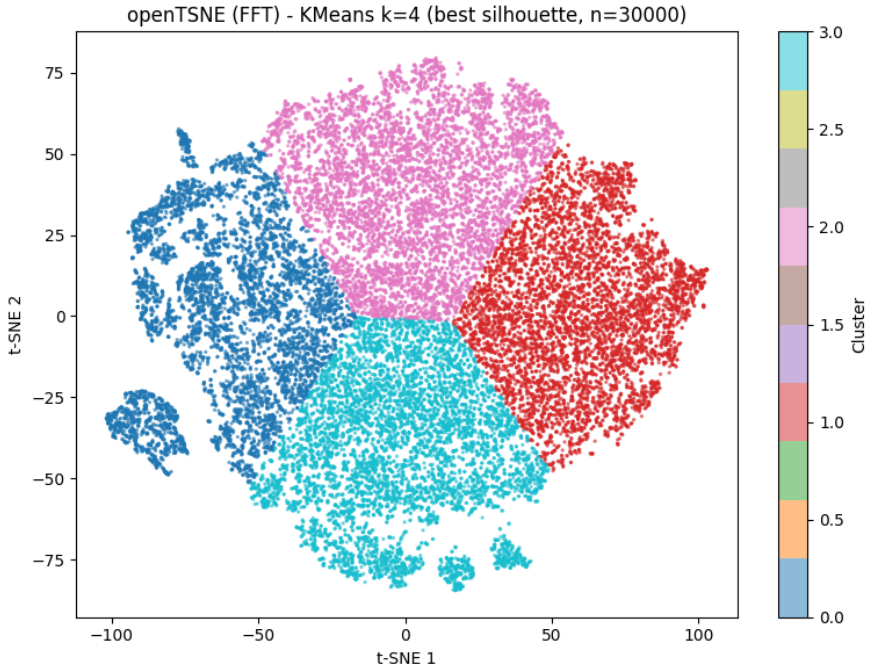
Next, we performed an experiment that first compared the clusters produced across the competing methods and ultimately determined which was most suitable for the task at hand. To visualize this, Figure 15 shows the clusters identified by the competing methods in t-SNE space. Upon inspection, it is clear that the different approaches are identifying very unique clusters, with especially HDBScan differing significantly from the other two methods. Next, the clustering methods were compared quantitatively based on their Silhoutte and NMI scores to determine which technique had the best performance. The outcome of this experiment is presented in Table 12. Overall, we see that K-Means significantly outperforms the other methods with it producing clusters that are more compact and better separated, while also agreeing more with the expert annotations. Furthermore, we see that HDBScan may not be appropriate for microglia clustering with it performing poorly on both metrics. One possible reason is that microglia exist along a broad morphological spectrum rather than in clearly discrete states [4], resulting in few sufficiently low-density regions for HDBSCAN to separate clusters along. Finally, while Agglomerative clustering did have a similar Silhoutte score to KMeans, its NMI was significantly lower likely due to it only identifying 2 clusters which were not able to capture the complexity of the true morphologies.



(a) Agglomerative clustering



(b) HDBSCAN clustering



(c) K-means clustering

Figure 15: Visualization of the hard clusters identified with different methods shown in t-SNE space

Clustering Method	Silhouette Score	NMI with manual annotations
K-Means	**0.388**	**0.2383**
HDBSCAN	0.249	0.0615
Agglomerative Clustering	0.342	0.0647

Table 12: Comparison of the clustering performance from the three competing methods. The Silhouette measures the cluster cohesion and separation while NMI quantifies how well the clusters agree with expert annotations

Based on their NMI scores alone, none of the clustering methods were able to represent the expert labels well. This can mainly be attributed to the use of 2D scans of 3D cells, meaning many human annotations were based on inferring missing parts of the staining that could indicate the cell type. For example, Figure 16 shows 2 homeostatic microglia, even though neither have many branches. However, because they have few starting points, they was deemed to be homeostatic, assuming parts of the branches were not captured in the staining. This makes it inherently difficult to achieve a high NMI score on this task, since some of the logic underlying the human annotations is visually not present in the scans.

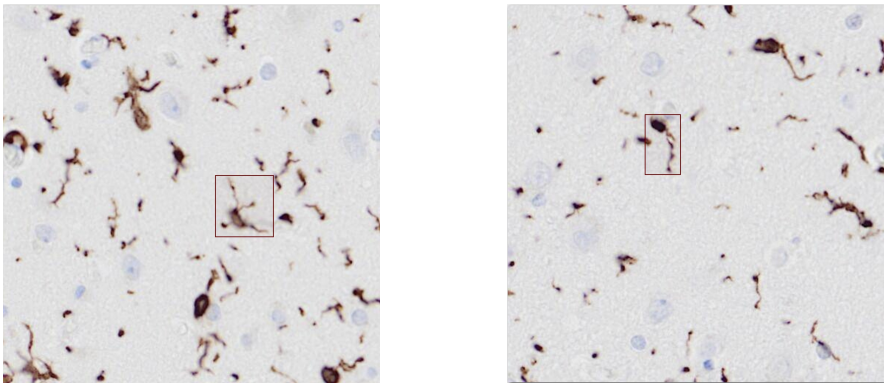


Figure 16: Two homeostatic microglia illustrating the ambiguity introduced by 2D staining of 3D cells. Despite neither having many branches, they were labeled as homeostatic due to having few starting points.

Next, Figure 18 enables us to visually understand the clusters identified by K-Means by showing the 4 closest cells to the centroids of each cluster. These can be seen as representative examples of the types of cells in each cluster. Notably, only some of the clusters identified relate to those found in previous literature [28] and in the manual annotations. With the aid of Figure 17, we can inspect this discrepancy by comparing how the cluster labels were assigned for the 97 manual annotations. Firstly, cluster 0 clearly represents the amoeboid morphology, with relatively few cells from other morphologies present. Further-

more, the homeostatic, hyper-ramified, and rod cells are more concentrated in clusters 2, 1, and 1, respectively. However, given the limited number of annotations, it is difficult to assess the model's performance on these morphologies. Nevertheless, even with only 49 annotations for the Reactive morphology, it is clear that it is spread across several clusters, indicating the morphological diversity of this cell state.

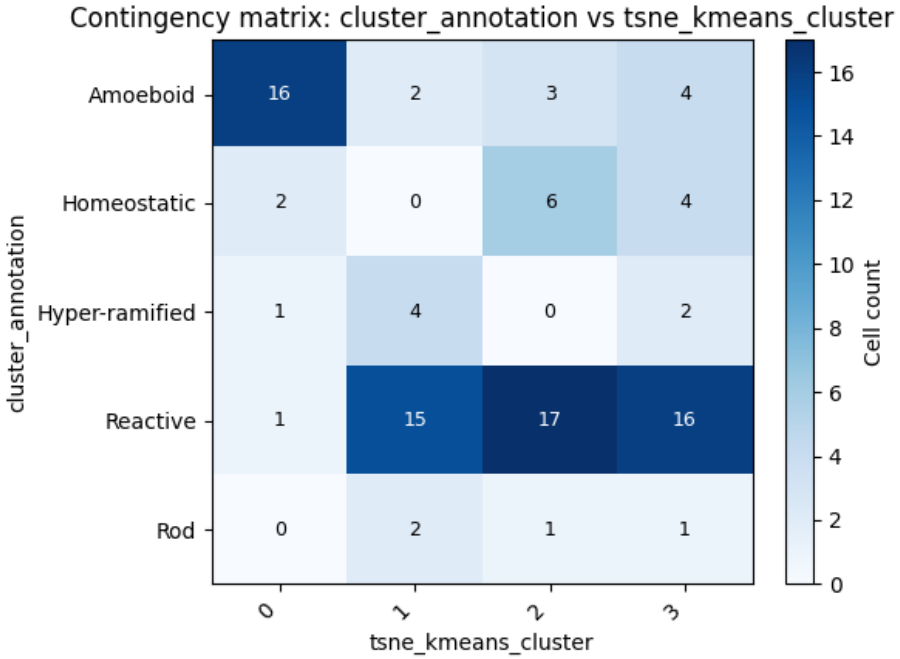
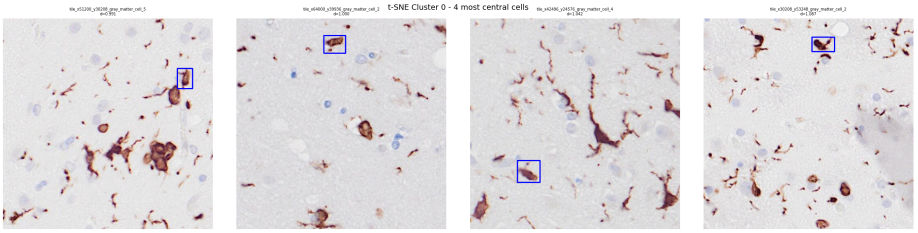
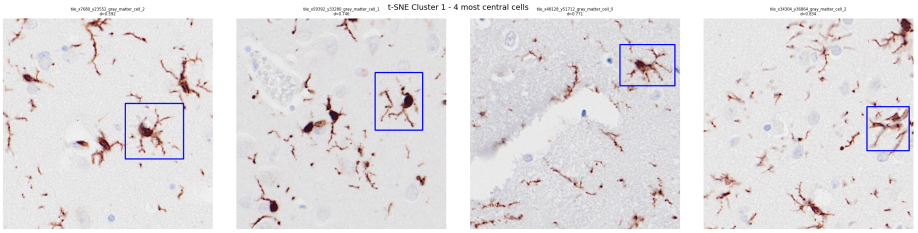


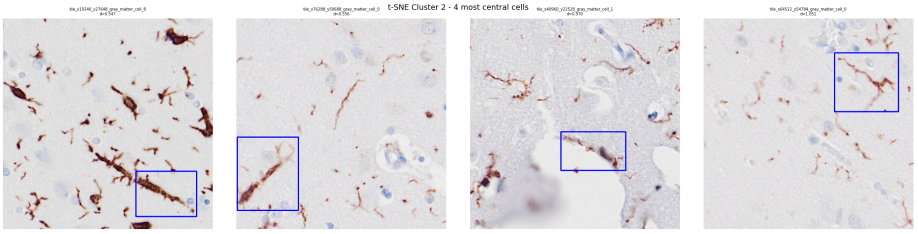
Figure 17: Contingency matrix showing how the 97 labeled cells (cluster_annotation) were divided across the clusters produced by KMeans on t-SNE.



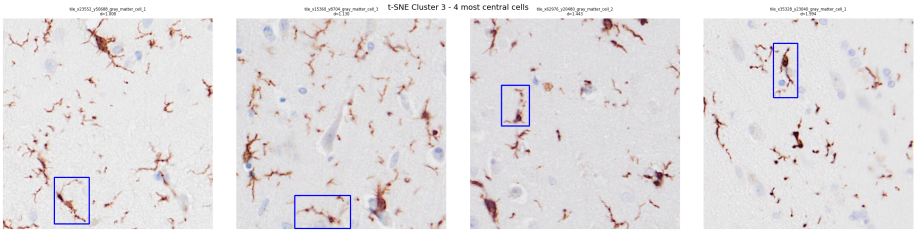
(a) Cluster 0



(b) Cluster 1



(c) Cluster 2



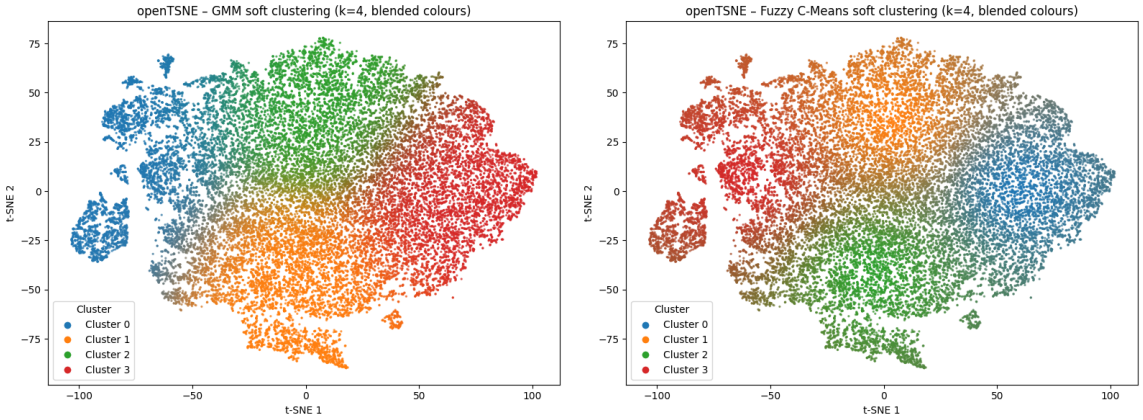
(d) Cluster 3

Figure 18: Each row shows representative cells for a different cluster, from Cluster 0 to Cluster 3. Representatives were selected as the cells closest to the centroids of the K-means clusters in the t-SNE embedding of morphological features.

7.4.2 Soft Clustering

To determine which soft clustering approach was most appropriate for microglia morphology representation, we compared their entropy and and percentage of assignments above 80%. GMM has a much lower mean entropy of 0.568 com-

pared to FCMeans at 1.20 indicating it is more concentrated in its cluster assignments. Furthermore, 66.4% of its cluster assignments are above 80% confidence compared to just 29.3% for FCMeans. With such confident predictions, GMM seems to not be fully capturing spectral aspect of microglia morphology and instead is assigning too confident predictions for many of the cells. Therefore, we have determined that FCMeans is more appropriate for the soft clustering of microglia. For a visual overview of the soft clusters produced by each method see Figure 19.



(a) GMM clustering (b) FCMeans clustering

Figure 19: Comparison of the soft clusters identified using GMM and FCMeans visualized in t-SNE space. The color gradient shows how the soft cluster assignments change with increasing distance from the centroids.

Overall, the main findings of our cluster experiments are the following. KMeans was deemed to be the best hard clustering method, with it achieving both a higher NMI and Silhouette score than the other competing methods. However, the clusters identified only partially align with the manual annotations, with them correctly identifying the amoeboid morphology but struggling with the others. Finally, the best soft clustering method was determined to be FCMeans as it captured the spectral aspect of the microglia cell states better than GMM.

7.5 Morphological Features Statistical Analysis

This section presents the results of the statistical tests conducted to assess whether there are differences between the three diagnosis groups in terms of their morphological features. As outlined in Section 5.7 the centenarians were divided into two subgroups based on pathology severity, with the AD patients making up the final group. The results presented in each subsection compare all three groups to each other in a pairwise manner.

7.5.1 Mixed Effect Modeling

As discussed in 5.7.1, a mixed effect modeling approach was used to determine if there are differences in individual morphological features across diagnosis groups when accounting for sex. Table 13 shows the results of this analysis, with a coefficient greater than 0 indicating higher values in diagnosis group 1 than in diagnosis group 2. Overall, very few individual features differ substantially across the diagnosis groups, with the main insight being the 100+_without_pathology group appears to have a more complex branching structures than the AD patients. This is indicated by their higher sholl_sum and lower sholl_min_radius which could represent more ramified microglia. However, this result should be interpreted with care, for example, while the sholl_min_radius has a relatively low p value at 0.0642, the q value when accounting for the false discovery rate is 0.321, indicating that this result is unlikely to be significant. Furthermore, as discussed in 7.2, our U-Net model has its highest errors on the sholl features, indicating that noise may be contributing to these results. See Table 6 for more information. As a whole, there are multiple possible reasons for the lack of significant differences in individual morphological features. For example, our dataset could be too small to capture differences that are actually present. Alternatively, it is also likely that these features in isolation do not capture sufficient morphological variation to distinguish cell states, and must instead be considered in combination with one another.

Diagnosis group 1	Diagnosis group 2	Feature	Coefficient	p-value	FDR q-value
AD	100+_with_pathology	cell_convex_circularity	-0.005441	0.376005	0.523926
AD	100+_with_pathology	num_components	-0.129700	0.419141	0.523926
AD	100+_with_pathology	sholl_min_radius	0.432173	0.175898	0.523926
AD	100+_with_pathology	sholl_sum	-10.645139	0.272536	0.523926
AD	100+_with_pathology	soma_perimeter	-0.971973	0.725214	0.725214
AD	100+_without_pathology	**sholl_min_radius**	**0.854853**	**0.064269**	**0.321347**
AD	100+_without_pathology	cell_convex_circularity	-0.004831	0.502739	0.844757
AD	100+_without_pathology	soma_perimeter	1.854314	0.506854	0.844757
AD	100+_without_pathology	num_components	-0.032994	0.869638	0.981871
AD	100+_without_pathology	**sholl_sum**	**0.288891**	**0.981871**	0.981871
100+_with_pathology	100+_without_pathology	num_components	0.099593	0.307884	0.513139
100+_with_pathology	100+_without_pathology	sholl_sum	6.335730	0.293620	0.513139
100+_with_pathology	100+_without_pathology	soma_perimeter	2.149428	0.167797	0.513139
100+_with_pathology	100+_without_pathology	sholl_min_radius	0.140780	0.432837	0.541046
100+_with_pathology	100+_without_pathology	cell_convex_circularity	0.000783	0.806862	0.806862

Table 13: Pairwise mixed-effects model results. The compared features were `cell_convex_circularity`, `num_components`, `sholl_min_radius`, `sholl_sum` and `soma_perimeter`. The coefficient represents the estimated difference in the feature value between diagnosis group 1 and diagnosis group 2 after adjusting for sex. Positive coefficients indicate higher values in diagnosis group 1, whereas negative coefficients indicate higher values in diagnosis group 2. The p value is the probability that diagnosis group 1 is greater than diagnosis group 2 and the q value represents the FDR rate as multiple features were tested.

7.5.2 PERMANOVA

As discussed in Section 5.7.2, a PERMANOVA analysis was performed to assess whether a linear combination of individual morphological features differ across the diagnosis groups. The selected features were the following: cell_convex_circularity, num_components, sholl_min_radius, sholl_sum and soma_perimeter. Across 5000 permutations there was between a 20.9% and 27.9% chance that the groups differ with one another on the selected features. See Table 14 for more details. Interestingly, none of the pairwise tests indicated statistically significant differences between the diagnosis groups. This result is consistent with the findings of our mixed effect analysis in Section 7.5.1, which suggest that although there may be morphological difference between the groups, these are not captured through tests on the morphological features themselves. This indicates that exceptional longevity is unlikely linked to specific morphological features or even linear combinations of them and instead there are more complex nonlinear relationships at play that only more sophisticated testing methods like our cluster Dirichlet analysis can detect. However, we can still gain insights from these results by determining which pairs of diagnosis groups are most similar to each other. For example, since the 100+_with_pathology and 100+_without_pathology groups

have the the lowest p value of the pairwise tests, these groups can be considered to be the most different. This finding is consistent with the results presented in 7.6.3, where we saw the only statistically significant cluster proportion difference on cluster 1. Furthermore, since AD and 100+_with_pathology have the highest p-value, these groups may be the most morphologically similar of all pairwise comparisons. This highlights that on these particular morphological features, 100+_with_pathology participants actually have more in common with AD patients than with their healthy counterparts.

Comparison	p_{perm}
`AD` vs `100+_with_pathology`	0.279344
`AD` vs `100+_without_pathology`	0.210558
`100+_with_pathology` vs `100+_without_pathology`	0.209558

Table 14: Results for pairwise PERMANOVA test comparing the diagnosis groups on a linear combination of morphological features. The p value represents the probability that the diagnosis groups differ. The compared features were `cell_convex_circularity`, `num_components`, `sholl_min_radius`, `sholl_sum` and `soma_perimeter`.

7.6 Cluster Proportions Statistical Analysis

This section presents the results of our statistical analysis, comparing whether the diagnosis groups differ in terms of cluster proportions and variance. These clusters represent the different cell states present within the scans. Therefore, by comparing their proportions, we can understand how cellular activity differed between diagnosis groups.

Prior to formal testing, the cluster proportions for each diagnosis group were visualized in Figure 20. Each point represents an individual, each vertex represents a cluster, and the color indicates the diagnostic group: red for AD, dark blue for 100+_with_pathology, and light blue for 100+_without_pathology. Overall, while there is a significant overlap between diagnosis groups, there are some visible differences between them, especially in cluster 3.

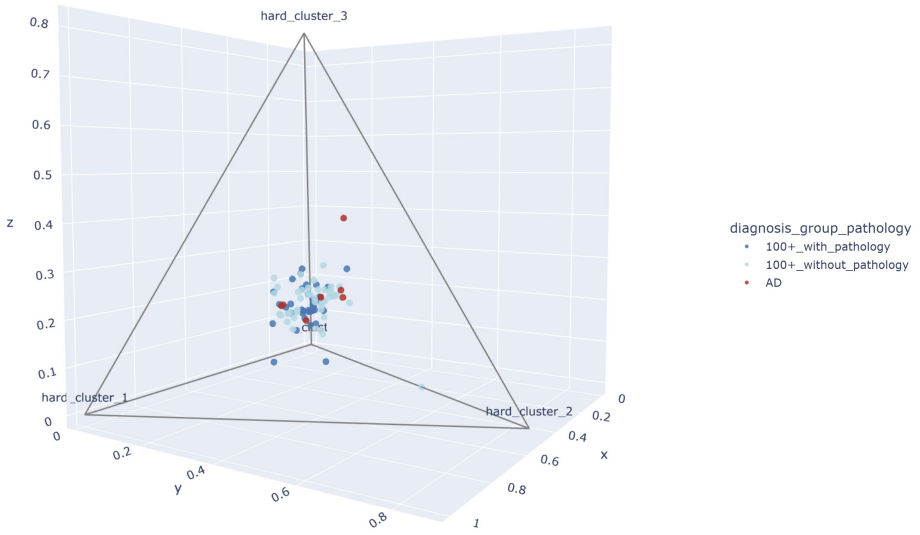


Figure 20: Visualization of the hard cluster proportions present in all individuals. Red points represent AD patients, dark blue points represent centenarians with pathology, and light blue points represent centenarians without pathology.

7.6.1 AD vs 100+_without_pathology Dirichlet

Firstly, we conducted an experiment to determine if the cluster proportions differed between the AD and 100+_without_pathology groups. By identifying and quantifying these differences we gain greater insights into the link between AD and microglia morphology. Table 15 shows the results of the Dirichlet analysis that compares the mean hard cluster proportions between the two groups. We can see that hard cluster 3 has the greatest difference between groups, with it being more common in the AD patients. Specifically, the results imply that the AD group has 7% more reactive microglia, as cluster 3 generally represents the reactive morphology. This finding is biologically plausible, since increased microglial activation can promote neuroinflammation, and chronic neuroinflammation has been linked to the development and progression of Alzheimer's disease [27]. See Section 2.1 for more information. Furthermore, this aligns with the findings presented in [33] who also identified a link between a similar morphology and AD progression. A complementary hypothesis is that, cluster 3 could represent cells that have undergone senescence with these being cells that remain metabolically active but are unable to function properly [23]. However, future research is necessary to confirm this by also including other biological data beyond cell morphology.

While the p value associated with this difference is relatively low at 0.071, it is not statistically significant. This can mainly be attributed to the small number of data points available, especially for the AD group with only 7 individuals in the cohort. Moreover, based on Figure 20 we can see that a few outliers may be responsible for the majority of this difference. This begs the question, is the increased presence of cluster 3 in AD patients a robust finding? To address this, we have implemented two additional robustness tests: Leave On Out Cross Validation (LOOCV) and the comparison of soft cluster proportions. Firstly, LOOCV allows us to assess the impact of outliers on our findings without removing them entirely from our dataset. This involved repeating the experiment 7 times each with a different AD individual left out of the dataset. For each data split we calculated the p value and considered the finding to be present in this split if the p value was less than 0.1. Figure 21 shows the results of this analysis. With only 28% of the cross validation splits resulting in a sufficiently low p value, we see that this finding may not be robust as it is influenced heavily by outliers. However, across all validation splits, the mean proportion of cluster 3 for AD is greater than for 100+_without_pathology, indicating that there could still be a difference but it is not statistically significant in this small dataset. Next, we assessed the robustness of this finding by performing the same analysis on the soft cluster proportions. These can be seen as smoother versions of the hard cluster proportions since for many of the patients, their proportions are more evenly spread across all clusters. Table 16 shows the results of this analysis with cluster 3 now being over 8% higher in AD scans, but this difference is much less statistically significant with a p value of 0.22. This shows that when accounting for the spectral nature of microglia morphologies there is still more of cluster 3 in AD patients. However, the substantially higher p-value observed in the soft cluster analysis suggests that the statistical significance of the hard clustering experiment may be inflated. This inflation could be attributed to cells residing on morphological boundaries being assigned to discrete clusters, when in reality they represent intermediate or transitional morphological states. Overall, both of our robustness tests indicated that our initial finding could indeed present although it is influenced by both outlier individuals and the oversimplification caused by hard clustering methods.

Cluster	Δ mean (AD $-$ 100+_without_pathology)	CI 2.5%	CI 97.5%	p(AD $>$ 100+_without_pathology)
hard_cluster_0	-0.029645	-0.103399	0.036173	0.273
hard_cluster_1	-0.005172	-0.105167	0.099065	0.456
hard_cluster_2	-0.036015	-0.108322	0.023380	0.221
hard_cluster_3	0.070832	-0.011361	0.164518	**0.929**

Table 15: Cluster mean differences between the AD and 100+_without_pathology groups based on the Dirichlet bootstrap analysis.

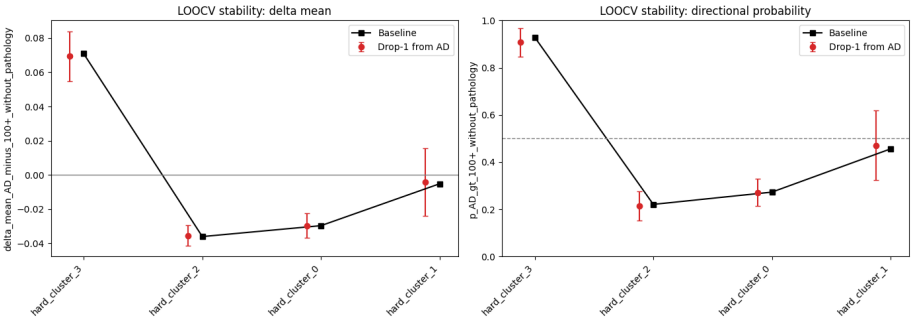


Figure 21: Robustness test of cluster mean differences between AD and 100+_without_pathology using LOOCV on AD patients.

Cluster	Δ mean (AD − 100+_without_pathology)	CI 2.5%	CI 97.5%	p(AD > 100+_without_pathology)
soft_cluster_0	0.007228	-0.025450	0.040562	0.673674
soft_cluster_1	-0.016172	-0.064580	0.032214	0.253253
soft_cluster_2	0.000861	-0.009721	0.010981	0.574575
soft_cluster_3	0.008082	-0.011861	0.028976	**0.781782**

Table 16: Soft cluster mean differences between the AD and 100+_without_pathology groups based on the Dirichlet bootstrap analysis.

Finally, we performed an experiment comparing the variance in cluster proportions between the diagnosis groups. This allows us to better understand if the groups are homogeneous or if there are many different microglial profiles present within them. Table 17 shows results of this experiment by analyzing the α values of each group's Dirichlet distributions. Overall, the variance of all cluster proportions is higher in the 100+_without_pathology group compared to AD, indicating greater morphological heterogeneity than in the AD group. This may reflect the fact that these individuals vary substantially from one another in factors such as sex, lifestyle, comorbidities, and age at death, with some individuals living to 111 years. It may also suggest that there are multiple microglial morphological profiles associated with living beyond 100 years of age, rather than a single characteristic profile. However, these results are not statistically significant which again could be attributed to the small dataset.

Cluster	$\Delta\text{Var}_{AD-100+\_wo\_p}$	**CI 2.5%**	**CI 97.5%**	$p(\text{AD} > 100+\_\text{without\_pathology})$
hard_cluster_0	-0.015731	-0.044711	0.005106	0.278
hard_cluster_1	-0.011710	-0.026369	0.003813	0.271
hard_cluster_2	-0.018316	-0.050862	0.005972	0.279
hard_cluster_3	-0.010644	-0.026136	0.006894	0.280

Table 17: Cluster variance differences between the AD and 100+_without_pathology groups based on the Dirichlet-implied bootstrap analysis.

7.6.2 AD vs 100+_with_pathology Dirichlet

As discussed in Section 6.5, we have conducted several experiments that compare the AD patients with centenarians that achieved an extraordinary age despite them having the AD pathology. By comparing these groups we can better understand the link between microglia morphology and an individual's health outcomes. Tables 18 and 19 show the difference in hard and soft cluster proportions for the AD vs 100+_with_pathology groups. As discussed in Section 7.6.1, the soft clustering results can be used to validate whether the hard clustering findings are robust to variation in microglia's spectral morphology. These groups differ on the greatest number of clusters compared to all other pairwise comparisons. While none of the differences are significant, they still hint that centenarians with the AD pathology have fairly different microglial profiles than those that died earlier with AD. More specifically, we see that AD patients have between 1.77% and 2.1% more of cluster 0, depending on comparing soft and hard clusters. Since cluster 0 represents fully activated (amoeboid microglia), this finding is biologically plausible, since increased microglial activation can promote neuroinflammation, which has been linked to the development and progression AD [27]. We also observe that AD patients have a lower proportion of cells from cluster 1 relative to centenarians with pathology. This is consistent with the finding above due to the fact that cluster 1 predominantly represents resting microglia, meaning a reduced proportion in AD patients reflects the same underlying shift toward a more activated microglial state. Finally, as with the previous pairwise comparison, AD patients had between 1.4% and 4.1% more cells from cluster 3 than centenarians with pathology. Since this difference is present across both centenarian comparisons, it may represent a morphological signature that is uniquely elevated in the AD group.

Cluster	Δ mean (AD − 100+_with_pathology)	CI 2.5%	CI 97.5%	p(AD > 100+_with_pathology)
hard_cluster_0	**0.020951**	-0.017364	0.054724	**0.881**
hard_cluster_1	**-0.064873**	-0.143130	0.029970	**0.083**
hard_cluster_2	0.002459	-0.026601	0.030297	0.599
hard_cluster_3	**0.041463**	-0.010663	0.107320	**0.918**

Table 18: Dirichlet mean differences between AD and 100+_with_pathology for hard clustering. Positive values indicate a higher estimated cluster proportion the AD group.

Cluster	Δ mean (AD − 100+_with_pathology)	CI 2.5%	CI 97.5%	p(AD > 100+_with_pathology)
soft_cluster_0	**0.017741**	-0.015856	0.048835	**0.853**
soft_cluster_1	**-0.032008**	-0.078782	0.021214	**0.115**
soft_cluster_2	-0.000227	-0.012189	0.010766	0.504
soft_cluster_3	**0.014495**	-0.008080	0.036994	**0.894**

Table 19: Dirichlet mean differences between AD and 100+_with_pathology for soft clustering. Positive values indicate a higher estimated cluster proportion in the AD group

Next, we conducted an experiment comparing the variance of cluster proportions across the diagnosis groups. This enables us to understand if there are specific microglial profiles associated with longevity and AD or or whether multiple distinct morphological profiles can lead to the same clinical outcomes. Table 20 shows the variance differences between the AD and 100+_with_pathology on hard cluster proportions. Interestingly, we see that the variance of all clusters proportions is higher in the AD group, which was not the case when comparing AD to 100+_without_pathology. This may indicate the the 100+_with_pathology group is more homogeneous than the AD and 100+_without_pathology groups, which hints towards a specific morphological profile associated with successfully handling AD into old age. Future research could build on this finding by first verifying this morphological profile on a larger dataset and then developing drugs that try to shift living AD patient's microglia to this ideal morphological profile.

Cluster	$\Delta\text{Var}_{AD-100+\_w\_p}$	**CI 2.5%**	**CI 97.5%**	p(AD > 100+_with_pathology)
hard_cluster_3	0.003180	-0.001464	0.008195	**0.833**
hard_cluster_2	0.002841	-0.001640	0.007111	**0.813**
hard_cluster_0	0.002460	-0.001247	0.006075	**0.835**
hard_cluster_1	0.001596	-0.001691	0.004692	**0.768**

Table 20: Dirichlet-implied cluster variance differences between AD and 100+_with_pathology. Positive values indicate greater variance in the AD group.

7.6.3 100+_with_pathology vs 100+_without_pathology Dirichlet

The final pairwise experiment we have conducted compares the two centenarian groups to assess whether the presence AD is linked to specific microglia morphologies in centenarians. Just like the previous pairwise analyses above, we first compared the differences in mean hard and soft cluster proportions across the diagnosis groups. The results of the hard cluster differences are presented in Table 21, while Table 22 shows the soft cluster results. Overall, we see that across both clustering methods the groups differ significantly, especially on clusters 0 and 1. For example, when using hard clusters, the centenarians with AD appear to have 6.18% more of cluster 1 with a p value of 0.027. This difference is also considerable when assessing the soft cluster proportions with this identifying a 1.7% increase with a still significant p value of 0.047. Biologically, cluster 1 represents resting, also known as homeostatic microglia, indicating that the centenarians with AD typically have more microglia surveying the environment compared to their healthy counterparts. This finding is further validated by the substantial differences in cluster 0 across the groups. Here, we saw that centenarians with the pathology had 5.1% less of hard cluster 0 compared to the healthy centenarians. This difference was also present in the soft clusters, although they indicate a 1.1% decrease with a p value of 0.099. Since cluster 0 represents the activated amoeboid microglia, we see that centenarians with AD have less microglia performing phagocytosis. Overall, the differences in cluster 0 and 1 proportions both point towards the same conclusion. This being that centenarians with AD pathology had fewer activated and more resting microglia compared to the centenarians without the pathology.

Cluster	Δ mean (with_pathology − without_pathology)	CI 2.5%	CI 97.5%	p(with_pathology > without_pathology)
hard_cluster_0	**-0.051907**	-0.115587	0.006204	**0.069**
hard_cluster_1	**0.061818**	-0.001213	0.125797	**0.973**
hard_cluster_2	-0.039674	-0.102619	0.012221	0.178
hard_cluster_3	0.029762	-0.031091	0.097552	0.696

Table 21: Dirichlet mean differences between 100+_with_pathology and 100+_without_pathology for hard clustering. Positive values indicate a higher estimated cluster proportion in the 100+_with_pathology group.

Cluster	Δ mean (with_pathology − without_pathology)	CI 2.5%	CI 97.5%	p(with_pathology > without_pathology)
soft_cluster_0	**-0.011227**	-0.027842	0.005637	**0.099**
soft_cluster_1	**0.017072**	-0.003101	0.035367	**0.953**
soft_cluster_2	0.000732	-0.006801	0.007837	0.572
soft_cluster_3	-0.006576	-0.017115	0.003556	0.116

Table 22: Dirichlet mean differences between 100+_with_pathology and 100+_without_pathology for soft clustering. Positive values indicate a higher estimated cluster proportion in the 100+_with_pathology group.

Building on our previous analyses, we compared the hard cluster variance across the groups to determine which centenarian cohort was more homogeneous. The results of this analysis can be found in Table 23. The main findings here agree with the theory presented in 7.6.2, namely that the 100+_with_pathology group has a much lower variance than the other diagnosis groups. For example, we see that it has a lower variance compared to the healthy centenarians on all clusters, which was also the case when comparing to the AD patients. This again leads us to believe that the 100+_with_pathology group has a specific morphological profile that is linked to their ability to combat AD well into their old age.

Cluster	$\Delta\text{Var}_{with\_pathology - without\_pathology}$	**CI 2.5%**	**CI 97.5%**	p(with_pathology > without_pathology)
hard_cluster_2	-0.021390	-0.051299	0.000925	0.121
hard_cluster_0	-0.018378	-0.044819	0.000597	0.108
hard_cluster_3	-0.014104	-0.027862	0.000656	0.098
hard_cluster_1	-0.013546	-0.027454	0.000987	0.149

Table 23: Dirichlet-implied cluster variance differences between 100+_with_pathology and 100+_without_pathology (bootstrap). Positive values indicate greater variance in the 100+_with_pathology group.

8 Conclusion

This thesis aimed to develop an end-to-end machine learning pipeline for microglia morphology analysis. This goal was exceeded by implementing and evaluating multiple competing methods at each stage of the pipeline and developing several novel microglia specific components that significantly improved performance. Together, the pipeline detected, segmented, and extracted the morphological features of over one million microglia, enabling a rigorous statistical analysis that linked specific morphological profiles to both AD and longevity.

There are several technical findings that this work uncovered. Firstly, human microglia detection performance can be improved by as much as 2% when pretraining on rat microglia. To the best of our knowledge, this is the first work to employ transfer learning of this nature to a microglia computer vision task, meaning it is a truly novel contribution. Next, we demonstrated that our custom active learning strategy, which priorities morphological diversity, lead to the greatest performance improvement with it beating both the edge confidence approach and additional random training data. Given this, our approach represents a novel contribution that future studies in cell detection could benefit from, extending beyond the microglia field. For example, in the detection of neurons or other cell types that exhibit high morphological diversity.

Regarding segmentation, we found that our custom U-Net model significantly outperformed SAM on all metrics due to it learning microglia specific structural features that are missing from a general segmentation model like SAM. Additionally, our robust performance analysis indicated that while our model is not able to predict all morphological features perfectly, it does exhibit strong overall performance with a morphology score of over 90%. Furthermore, this custom morphology score proved to capture truly new insights that would have otherwise been missed using traditional metrics. For example, it enabled us to determine that postprocessing improved our morphological performance by over 8% while the Dice and IoU showed little change. Beyond this, our segmentation was further improved through domain-specific augmentations and the inclusion of topology-aware loss components. This again showed that the custom microglia specific components implemented in this work enabled our models to achieve better performance than would be possible with a generic, domain agnostic approach.

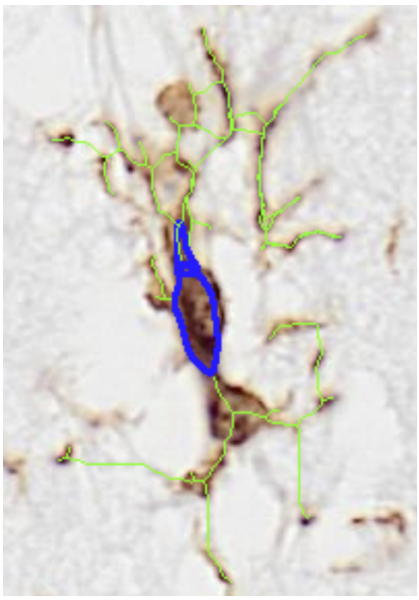
In our downstream analyses we found that t-SNE was a more suitable dimensionality reduction technique for microglia morphology, indicating the the features contain non-linear relationships. Following this, we identified K-means as the best hard clustering model for microglia due to it achieving the highest silhouette score and greatest agreement with manual the manual annotations. In addition, F-CMeans was concluded to be more appropriate than GMM for our task as it better modeled the spectral aspect of microglia morphology.

The biological findings from this work are numerous. Firstly, we found that statistical tests that operate directly on the morphological features, like our

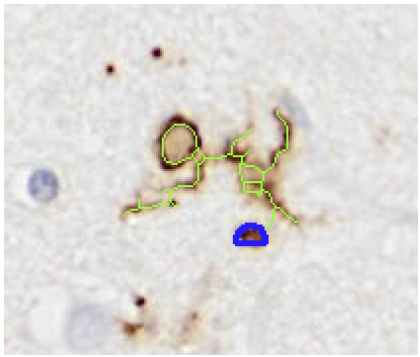
PERMANOVA and mixed effect models, are unlikely to have identified significant morphological differences between the groups. This indicates that if morphological differences are present, they are based on complex relationships between many morphological features. However, we were still able to gain insights from our PERMANOVA analysis by observing which groups were the most morphologically similar to one another. For example, on these specific morphological features, AD and 100+_with_pathology are the most similar diagnosis groups. Next, based on our Dirichlet analysis of both hard and soft cluster proportions, we found several morphological differences between the diagnosis groups. For example, cluster 3 appears to be more present in the AD patients than in either of the other diagnosis groups. These cells represent mostly reactive microglia that could be causing excessive neuroinflammation or perhaps have undergone senescence meaning they are no longer functioning properly. Next, the 100+_with_pathology group appears to have less amoeboid and more ramified microglia than both of the other groups. This indicates that their cells were performing phagocytosis at a lower rate, which may be linked to their exceptional longevity despite having the AD pathology. Furthermore, this difference is statistically significant between the centenarian groups ($p = 0.027$), indicating that separate morphological profiles are linked to longevity depending on whether the individual had AD, rather than a single profile shared across all centenarians. In addition to this, the 100+_with_pathology group had the lowest variance in their morphological profiles, indicating that there may be a specific profile that is linked to their longevity. This is in contrast to the AD and 100+_without_pathology groups which have a much higher variance, indicating that many morphological profiles can be present in these groups.

There are several limitations present in the current pipeline. Firstly, as is common with many YOLO implementations, our models struggled to identify touching microglia as separate and this error is propagated throughout the rest of the pipeline. Figure 22a shows an example of how two cells can incorrectly be treated as one at the end of the pipeline, due to a mistake upstream in the YOLO output. Next, regarding segmentation, since our model was only trained on 1009 labeled cells, it occasionally struggled to segment hollow somas — those in which only the outline is visible due to incomplete staining. Figure 22b shows an example of a strong overall cell segmentation but poor hollow soma detection. Beyond this, although our custom morphology score is a truly novel contribution, it may have a slightly circular nature to it. More specifically, to get our custom weights, we needed to start with an initial U-Net segmentation model that may have had bias in its performance which then impacted the features that were identified as being important for clustering. For example, if this initial model could not properly predict the Sholl min radius, then it will never be considered as an important clustering feature due to the excessive noise. Additionally, our evaluation of soft clustering may have been overly simplistic, as it favored models that had a high mean entropy. However, this may not fully describe a good soft clustering method, as a high entropy can also be artificially achieved without necessarily capturing the morphologi-

cal spectrum appropriately. Finally, while the statistical analysis was thorough, with such a small number of AD patients, we could only hint at subtle differences between the diagnosis groups rather than identify statistically significant patterns. Nonetheless, this work has identified some promising links between microglia morphology and longevity that future work can explore further with a larger dataset.



(a) Example of how the pipeline poorly handles touching microglia



(b) Incorrect skeletonization of touching microglia

Figure 22: Example microglias showing the limitations of our pipeline

The possibilities for future work are varied. Firstly, with such strong performance improvements in detection by pretraining on rat data, one could also apply a similar strategy to the segmentation task by also pretraining on other species data before finetuning on humans. Next, since the topological loss components were shown to improve segmentation performance, future work could make a custom loss function aimed specifically at predicting morphological features relevant for microglia. Finally, as previously mentioned, our current detection model sometimes struggled with multiple touching microglia. One possible solution that future work could implement would be a soma-aware non maximum suppression (NMS) step. In this custom component, NMS would take into account the different soma locations before discarding bounding boxes and thus be more likely to identify the two distinct microglia. Alternatively, future work could combine the YOLO bounding boxes and then do instance segmentation to identify the individual overlapping microglia.

In summary, this work has significantly contributed to the field of automatic microglia morphology analysis through the development of an end-to-end machine learning pipeline that achieves strong results in every stage. It has done so through several novel microglia specific contributions that have enabled us to achieve better performance than with a task-agnostic alternative. These include, but are not limited to, developing a custom active learning method, implementing microglia specific data augmentations as well as loss components and applying soft clustering to capture the spectral aspect of microglia morphologies. In addition to these technical contributions, we also furthered the scientific community's biological understanding of microglia's role in AD and longevity by identifying specific morphological profiles that are linked to the different diagnosis groups.

9 Reflection

This work did not achieve its strong results in a vacuum. Instead, it was enabled by the findings from several previous studies that motivated the choices implemented. For example, due to the results in [2], YOLO was the only detection architecture implemented in this pipeline, as the authors found it outperformed image processing techniques as well as other detection architectures. Furthermore, based on the authors in [1] separating their segmentation evaluation between soma and branches, we were inspired to adopt the same approach and think critically about what truly indicates strong segmentation performance for microglia. This deeper consideration on evaluation methodology ultimately led to the development of our own custom morphology metric. In addition to leveraging existing findings, this work has also contributed significantly to the field of automatic microglia morphology analysis through the development of a complete end-to-end ML pipeline as well as several custom microglia specific components. See Section 3 for more information on how this thesis relates to current research and which shortcomings it addresses.

Zooming out, this thesis sits within a broader set of trends and challenges shaping the Data Science and Artificial Intelligence Technology (DSAIT) field. Firstly, since the 2012 breakthrough of AlexNet [15], there has been a monumental shift in the field of computer vision, applying CNNs over traditional ML or image processing techniques for a variety of tasks. For example, they have been shown to perform incredibly well on image classification, object detection and semantic segmentation across a broad range of domains, which strongly influenced the choice to apply them in this thesis [45]. Moreover, this research actively addresses one of the main challenges affecting the field of microglia morphology analysis and data intensive tasks more broadly, with this being the large data annotation burden placed on experts. While previous literature has needed as many as 80,000 annotated cells to achieve strong results [11], we have shown that competitive performance is achievable with far fewer annotations

through inter-species transfer learning and our custom active learning method. This can also be seen as part of a larger trend, with recent research applying both active and transfer learning to a variety of domains with great success [18], [3]. Complementary to these broader trends, this thesis also contributes to the growing body of work applying YOLO and U-Net to the detection and segmentation of microscopy images, a development that has gained considerable momentum in recent years ([31], [1], [38], [44], [2]). This trend has unlocked new biological analyses that were previously infeasible due to the limitations of manual labeling and insufficient model accuracy. The impact of this is monumental, with new biological findings being discovered at a much greater pace than possible with previous techniques.

As discussed throughout the report and especially in Section 8, this pipeline has several limitations that impact its findings greatly. For example, the low sample size introduces the risk that the reported differences between diagnosis groups are not actually present within the larger population. Consequently, if future AD treatments were developed based on these findings alone, it is possible that a patient receives treatment that actually harms them. Furthermore, if I were a patient and was informed that the treatment I am receiving was based only on 7 AD data points, I would not feel safe taking it. Additionally, there is a separate risk introduced by the performance of the pipeline as a whole. As discussed in Section 7, while we do achieve strong results on all pipeline stages, it is not perfect. Furthermore, due to our pipeline design, errors can be propagated and compounded on through each stage of the pipeline, sometimes resulting in incorrect results for individual cells. These errors could lead to incorrect conclusions in downstream analyses, especially if there is bias in our models. For example, if they make more errors on ramified cells then other simpler morphologies. The impact of these errors could be significant and ultimately result in us reporting on morphological differences that are not present or alternatively, missing key ones that are. In the short term, such incorrect findings introduced into the scientific literature could impede our collective understanding of AD, longevity, and the role of microglia, ultimately delaying the development of effective treatments. While this would negatively affect patients, their families and medical professionals, it could also actively hinder other researchers who build upon this work, potentially causing a cascade of misdirected research efforts and wasted resources. A final stakeholder to consider when examining the real world risks of this work are the centenarians, AD patients and their families. Given that these findings rely on post-mortem brain tissue generously provided by these donors, there is an ethical obligation to report all derived insights with utmost responsibility. This necessitates cautious interpretation to ensure their contributions are not misrepresented through exaggerated claims or the overstatement of results. Overall, while these risks are present, we believe that all reasonable steps have been taken to mitigate them through the development of a robust evaluation framework, a detailed description of our experimental setup to ensure reproducibility, and by transparently acknowledging the limitations of our findings to ensure they are not overstated or misinterpreted.

10 Use of generative AI

Generative AI was used in several ways throughout this thesis, including on the pipeline as well as on the final report. Generative AI has not been used to replace critical thinking or personal engagement; instead all critical analysis, reasoning, and conclusions have been done by the author. The specific generative AI tools used were Claude Sonnet 4.5, 4.6 and Claude Code as well as GPT 5.2, 5.3 and 5.4. An overview of how generative AI was used can be found below.

Pipeline Development
Generative AI tools were used for brainstorming, to find academic sources and to write specific boilerplate code. Firstly, regarding brainstorming, generative AI was occasionally used to critique the pipeline and point weaknesses in the current experiments. For example, based on one weak point it identified, I decided to include the random additional training data as a baseline when comparing active learning strategies. It is important to note that all suggestions and feedback provided were thoroughly investigated, with generative AI mostly acting as a sparring partner. Next, generative AI was used to help find some academic sources. More specifically, I initially performed my own thorough literature review by finding sources using Google Scholar and the snowballing method, after which I used generative AI to search for any papers it felt were related but that were missing from my initial search. I then read the suggested papers in full to verify their relevance, some of which proved to be useful sources that had been absent from my initial search. Finally, generative AI was used to generate some boilerplate code, for example the bash scripts that ran my pipeline on DAIC with different parameters. All code was personally verified to be correct and functioning as intended.

Report
Firstly, generative AI was used to improve the writing quality and flow of the report including identifying grammatical errors and suggesting alternative vocabulary. For example, by prompting it with queries such as: 'What word could I use instead of 'Next' in this sentence?'. In addition to this, early drafts of the report were critiqued using Claude Code which helped determine if all stages of the pipeline were fully explained. Furthermore, generative AI tools were used to format the LaTeX tables and formulas in this report. However, results and formulas were manually checked to be accurate.

Presentation
No generative AI was used in the final presentation or any intermediate presentation throughout this Master's thesis.

11 Appendix

11.1 Morphological Feature Visualizations

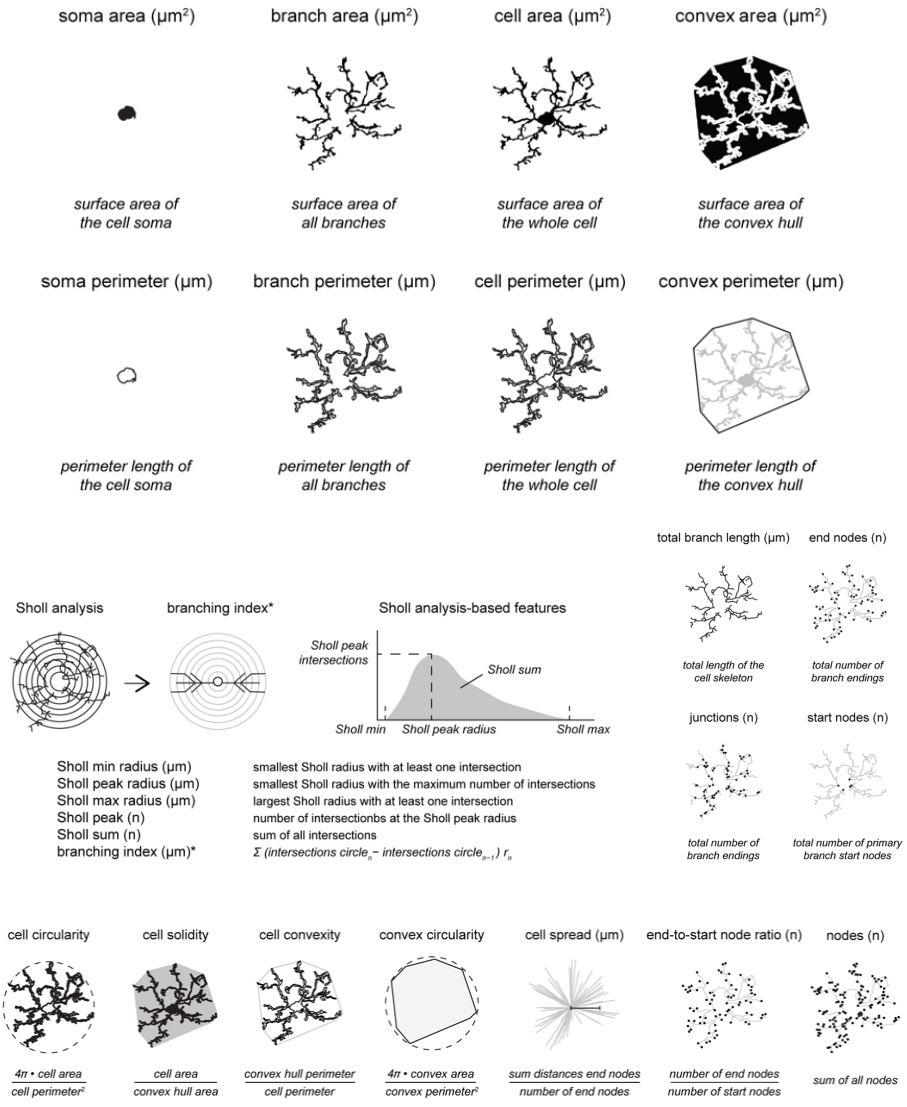


Figure 23: Overview of morphological features used, including cell features, Sholl analysis features, skeleton-based features, and derived ratios. All figures were found in [42]

.

References

[1] MohammadAmin Alamalhoda, Arsalan Firoozi, Alessandro Venturino, and Sandra Siegert. traice3d: A prompt-driven transformer based u-net for semantic segmentation of microglial cells from large-scale 3d microscopy images. *arXiv preprint arXiv:2507.22635*, 2025.

[2] Danish M. Anwer, Francesco Gubinelli, Yunus A. Kurt, Livija Sarauskyte, Febe Jacobs, Chiara Venuti, Ivette M. Sandoval, Yiyi Yang, Jennifer Stancati, Martina Mazzocchi, Edoardo Brandi, Gerard O'Keeffe, Kathy Steece-Collier, Jia-Yi Li, Tomas Deierborg, Fredric P. Manfredsson, Marcus Davidsson, and Andreas Heuer. A comparison of machine learning approaches for the quantification of microglial cells in the brain of mice, rats and non-human primates. *PLoS One*, 18(5):e0284480, 2023.

[3] Suat Atasever, Nuh Azginoglu, Dakun Satilmis Terzi, and Reyat Terzi. A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. *Clinical Imaging*, 94:18–41, February 2023.

[4] S. B. Beynon and F. R. Walker. Microglial activation in the injured and healthy brain: What are we really talking about? practical and theoretical issues associated with the measurement of changes in microglial morphology. *Neuroscience*, 225:162–171, 2012.

[5] Vanshu Bhardwaj, Sneha Kumari, Rishika Dhapola, Prajjwal Sharma, Samir Kumar Beura, Sunil Kumar Singh, Balachandar Vellingiri, and Dibbanti HariKrishnaReddy. Shedding light on microglial dysregulation in Alzheimer's disease: exploring molecular mechanisms and therapeutic avenues. *Inflammopharmacology*, 33:679–702, 2025.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. See also Section 9.3.2.

[7] Robert M. De Jager, Annie J. Lee, Alina Sigalov, and Mariko Taga. An image segmentation pipeline optimized for human microglia uncovers sources of morphological diversity in alzheimer's disease. *bioRxiv*, 2024.

[8] Yinglei Duan, Chongfang Han, Hua Zheng, Jing Yu, and Min Luo. Global, regional, and national burden of Alzheimer's disease and other dementias from 1990 to 2021: findings from the Global Burden of Disease Study 2021. *Frontiers in Aging Neuroscience*, 17:1678212, 2025.

[9] Gregory Gundersen. Dirichlet-multinomial. https://gregorygundersen.com/blog/2020/12/24/dirichlet-multinomial/, 2020. Accessed: 2026-04-19.

[10] Henne Holstege, Nina Beker, Tjitske Dijkstra, Karlijn Pieterse, Elizabeth Wemmenhove, Kimja Schouten, Linette Thiessens, Debbie Horsten, Sterre Rechtuijt, Sietske Sikkes, Frans W A van Poppel, Hanne Meijers-Heijboer,

Marc Hulsman, and Philip Scheltens. The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. *European Journal of Epidemiology*, 33(12):1229–1249, 2018.

[11] Chao-Hsiung Hsu, Yi-Yu Hsu, Be-Ming Chang, Katherine Raffensperger, Micah Kadden, Hoai T. Ton, Essiet-Adidiong Ette, Stephen Lin, Janiya Brooks, Mark W. Burke, Yih-Jing Lee, Paul C. Wang, Michael Shoykhet, Tsang-Wei Tu, et al. Stainai: quantitative mapping of stained microglia and insights into brain-wide neuroinflammation and therapeutic effects in cardiac arrest. *Communications Biology*, 8:462, 2025.

[12] Bradley T. Hyman, Creighton H. Phelps, Thomas G. Beach, Eileen H. Bigio, Nigel J. Cairns, Maria C. Carrillo, Dennis W. Dickson, Charles Duyckaerts, Matthew P. Frosch, Eliezer Masliah, Suzanne S. Mirra, Peter T. Nelson, Julie A. Schneider, Dietmar R. Thal, Bill Thies, John Q. Trojanowski, Harry V. Vinters, and Thomas J. Montine. National institute on aging–alzheimer's association guidelines for the neuropathologic assessment of alzheimer's disease. *Alzheimer's & Dementia*, 8(1):1–13, 2012.

[13] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[16] L. J. Lawson, V. H. Perry, P. Dri, and S. Gordon. Heterogeneity in the distribution and morphology of microglia in the normal adult mouse brain. *Neuroscience*, 39(1):151–170, 1990.

[17] Jake Lever, Martin Krzywinski, and Naomi Altman. Principal component analysis. *Nature Methods*, 14:641–642, 2017.

[18] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[19] Qingyun Li and Ben A. Barres. Microglia and macrophages in brain homeostasis and disease. *Nature Reviews Immunology*, 18(4):225–242, 2018.

[20] Leo Yu-Feng Liu, Yufeng Liu, and Hongtu Zhu. Masked convolutional neural network for supervised learning problems. *Stat*, 9(1):e290, 2020.

[21] Jacqueline M. Long and David M. Holtzman. Alzheimer disease: An update on pathobiology and treatment strategies. *Cell*, 179(2):312–339, 2019.

[22] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 4765–4774, 2017.

[23] Antonio Malvaso, Alberto Gatti, Giulia Negro, Chiara Calatozzolo, Valentina Medici, and Tino Emanuele Poloni. Microglial senescence and activation in healthy aging and alzheimer's disease: Systematic review and neuropathological scoring. *Cells*, 12(24):2824, 2023.

[24] Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

[25] Rosa C. Paolicelli, Giulia Bolasco, Fabio Pagani, Laura Maggi, Matteo Scianni, Patrizia Panzanelli, Marco Giustetto, Tiago A. Ferreira, Elisa Guiducci, Laure Dumas, Daniela Ragozzino, and Cornelius T. Gross. Synaptic pruning by microglia is necessary for normal brain development. *Science*, 333(6048):1456–1458, 2011.

[26] Marco Prinz, Steffen Jung, and Josef Priller. Microglia biology: One century of evolving concepts. *Cell*, 179:292–311, 2019.

[27] Richard M. Ransohoff. How neuroinflammation contributes to neurodegeneration. *Science*, 353(6301):777–783, 2016.

[28] Jack Reddaway, Peter Eulalio Richardson, Ryan J. Bevan, Jessica Stoneman, and Marco Palombo. Microglial morphometric analysis: so many options, so little consistency. *Frontiers in Neuroinformatics*, 17:1211188, 2023.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[31] Enes Sengun, Sharifa Alduraibi, et al. State-of-the-art deep learning methods for microscopic image segmentation: Applications to cells, nuclei, and tissues. *Journal of Imaging*, 10(12):311, 2024.

[32] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, Department of Computer Sciences, 2009.

[33] Ryan K. Shahidehpour, Rebecca E. Higdon, Nicole G. Crawford, Janna H. Neltner, Eseosa T. Ighodaro, Ela Patel, Douglas Price, Peter T. Nelson, and Adam D. Bachstetter. Dystrophic microglia are associated with neurodegenerative disease and not healthy aging in the human brain. *Neurobiology of Aging*, 99:19–27, 2021.

[34] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice – A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16560–16569, 2021.

[35] D. A. Sholl. Dendritic organization in the neurons of the visual and motor cortices of the cat. *Journal of Anatomy*, 87(4):387–406, 1953.

[36] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[37] Amanda Sierra, Jose M. Encinas, Juan J. Deudero, Jesse H. Chancey, Grigori Enikolopov, Linda S. Overstreet-Wadiche, Stella E. Tsirka, and Mustafa Maletic-Savatic. Microglia shape adult hippocampal neurogenesis through apoptosis-coupled phagocytosis. *Cell Stem Cell*, 7(4):483–495, 2010.

[38] Ilida Suleymanova, Dmitrii Bychkov, and Jaakko Kopra. A deep convolutional neural network for efficient microglia detection. *Scientific Reports*, 13(1):11139, 2023.

[39] Luke Sumberg, Rina Berman, Antoni Pazgier, Joaquin Torres, Jennifer Qiu, Bodhi Tran, Shannen Greene, Rose Atwood, Martin Boese, and Kwang Choi. A semi-automated and unbiased microglia morphology analysis following mild traumatic brain injury in rats. *International Journal of Molecular Sciences*, 26(17):8149, 2025.

[40] Illia Tsiporenko, Pavel Chizhov, and Dmytro Fishman. Going beyond u-net: Assessing vision transformers for semantic segmentation in microscopy image analysis. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi, editors, *Computer Vision – ECCV 2024 Workshops*, volume 15638 of *Lecture Notes in Computer Science*, pages 222–238. Springer, Cham, 2025.

[41] N. A. Valous, B. Lahrmann, W. Zhou, R. Veltkamp, and N. Grabe. Multistage histopathological image segmentation of iba1-stained murine microglias in a focal ischemia model: methodological workflow and expert validation. *Journal of Neuroscience Methods*, 213(2):250–262, 2013.

[42] Hilmar R. J. van Weering, Tjalling W. Nijboer, Maaike L. Brummer, Erik W. G. M. Boddeke, and Bart J. L. Eggen. Microglia morphotyping in the

adult mouse cns using hierarchical clustering on principal components reveals regional heterogeneity but no sexual dimorphism. *Glia*, 71(10):2356–2371, 2023.

[43] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[44] Alexander Zähringer, Janaki Manoja Vinnakota, Tobias Wertheimer, Philipp Saalfrank, Marie Follo, Florian Ingelfinger, and Robert Zeiser. Aistain: Enhancing microglial phagocytosis analysis through deep learning. *Cell Reports Methods*, 5(11):101207, 2025.

[45] Xia Zhao, Limin Wang, Yuhan Zhang, Xinbo Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99, 2024.